Int. J. Fuzzy Syst.
https://doi.org/10.1007/s40815-022-01324-w

Transparent but Accurate Evolutionary Regression Combining New Linguistic Fuzzy Grammar and a Novel Interpretable Linear Extension

Carmen Biedma-Rdguez[1] · María José Gacto[2] · Augusto Anguita-Ruiz[3] · Jesús Alcalá-Fdez[1] · Rafael Alcalá[1]

Received: 25 June 2021 / Revised: 18 November 2021 / Accepted: 8 April 2022

Abstract Scientists must understand what machines do (systems should not behave like a black box), because in many cases how they predict is more important than what they predict. In this work, we propose a new extension of the fuzzy linguistic grammar and a mainly novel interpretable linear extension for regression problems, together with an enhanced new linguistic tree-based evolutionary multiobjective learning approach. This allows the general behavior of the data covered, as well as their specific variability, to be expressed as a single rule. In order to ensure the highest transparency and accuracy values, this learning process maximizes two widely accepted semantic metrics and also minimizes both the number of rules and the model mean squared error. The results obtained in 23 regression datasets show the effectiveness of the proposed method by applying statistical tests to the said metrics, which cover the different aspects of the interpretability of linguistic fuzzy models. This learning process has obtained the preservation of high-level semantics and less than 5 rules on average, while it still clearly outperforms some of the previous state-of-the-art linguistic fuzzy regression methods for learning interpretable regression linguistic fuzzy systems, and even to a competitive, pure accuracy-oriented linguistic learning approach. Finally, we analyze a case study in a real problem related to childhood obesity, and a real expert carries out the analysis shown.

Keywords Regression · Linguistic modeling · Evolutionary fuzzy Systems · eXplainable Artificial Intelligence (XAI) · Interpretability · Transparency

✉ Rafael Alcalá
alcala@decsai.ugr.es

Carmen Biedma-Rdguez
biedmardguez@gmail.com

María José Gacto
mjgacto@ugr.es

Augusto Anguita-Ruiz
augustoanguitaruiz@gmail.com

Jesús Alcalá-Fdez
jalcala@decsai.ugr.es

[1] Department of Computer Science and Artificial Intelligence, DaSCI, University of Granada, Granada, Spain

[2] Department of Software Engineering, DaSCI, University of Granada, Granada, Spain

[3] Department of Biochemistry and Molecular Biology II, Institute of Nutrition and Food Technology "José Mataix," Center of Biomedical Research, University of Granada, Granada, Spain

1 Introduction

The era of Big Data, Deep Learning (DL) and the Internet of Things (IoT), has been a breakthrough for Artificial Intelligence (AI), making it one of the most revolutionary technologies to date. However, the tremendous advances that AI has experienced in recent years have caused a wave of concern, since in most cases we do not know how the software learns and makes decisions. While the term of *eXplainable Artificial Intelligence* (XAI) is relatively new, the problem of explaining AI techniques actually became a challenge many years ago. In fact, as stated in[1] [1], it dates back to 1991, when Dr. Pomerleau studied how a neural network thinks in an attempt to explain why an autonomous car decided to leave it's lane on a bridge after thousands of tests. More recent cases like the challenge of

[1] D. Castelvecchi, Can we open the black box of AI? (2016) Nature.

Published online: 12 July 2022

relying on autonomous systems for military operations regarding weapons [2] or problems involving dilemmas of life and death decisions, such as the most recent IBM Watson at the National Hospital of Denmark [3], which made a very serious mistake in recommending "deadly treatment" for cancer patients, are generating quite a bit of controversy regarding a real need to explain AI. One such example is the case of the well-known DL that cannot explain how it makes its decisions despite its impressive results. This is known as the "black box problem," which has been recently discussed in the prestigious journal Nature [1]. Today, it is an open problem that people are being forced to work on at a fast pace [4].

For this reason, and in order to help AI use become a widespread reality, researchers are taking into account all aspects related to ethics [5], Law [6], and technology [7]. Recently, ACM issued a Declaration on "Algorithmic Transparency and Accountability," which establishes a set of principles that are consistent with the ACM Code of Ethics to support the benefits of algorithmic decision-making while addressing ethical and legal issues [8]. Among such principles, the explanation is identified as relevant. In addition, a new European General Data Protection Regulation (GDPR3) was applied on 25 May 2018 [6, 9], which replaced the previous 20-year-old Directive 95/46/EC. GDPR3 is concerned with transparency and protection of the nature of people when personal data are processed freely. For the first time, a highly discussed form of a "right to explanation" is inferred by some experts [6] based on the regulations for the automated decision-making from Article 22, Recital 71: "... the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached ...". In terms of technological issues, the issue of explainability in AI is also highlighted in the challenge established by the Defense Advanced Research Projects Agency of the United States (DARPA) [10]: "Although current AI systems offer many benefits in many applications, their effectiveness is limited by the lack of explanations when interacting with humans." Overall, all these give rise to what has recently become known as XAI [10–13].

Machine learning (ML) is becoming ubiquitous in both basic research and industry. Consequently, non-expert users, that is, users without a strong base in AI, require a new generation of explainable AI systems. Scientists must also understand what machines do [1] (systems should not behave like a black box), because in many cases how they predict is more important than what they predict. These systems should be directly interpretable, i.e., they should explain their behavior in a way that they can be understood [14, 15]. Clear examples can be found in problems in healthcare [16] and in particular in bio-medicine [17], banking advice, insurance [18], legal decision-making, robotics, planning, and many others. Therefore, there is a need to continue investigating the improvement of machine learning techniques with an inherent high explanatory power, as the most appropriate alternative "for high-stakes prediction applications that deeply impact human lives" [14].

Recently, Barredo et. al. [13] have stated that "an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand." Based on the revised proposals and bibliographic studies, in this contribution, the authors also propose two taxonomies of XAI techniques and clarify the real meaning of the different concepts usually used in the context of XAI. Interpretability is initially defined as "a passive characteristic of a model referring to the level at which a given model makes sense for a human observer" (which is "also expressed as transparency"). However, explainability is referred to "as an active characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions." Based on this, they set a first main division in their categorization [13]:

- Transparent ML models: Models that can be understood by themselves. Some examples are Logistic / Linear Regression, Decision Trees, Rule-Based Learners, Fuzzy Logic-Based models, etc.
- Post hoc explainability techniques for ML models: Those that need a posterior engine or process to explain the model decisions when they cannot be considered to be transparent.

Furthermore, they also introduce the concept of "simulatability" as the most interpretable level within the transparent ML models. According to their definition, it "denotes the ability of a model to be simulated or thought about strictly by a human, hence complexity takes a dominant place in this class." For example, Linear Regression must make use of human-readable predictors with a minimum of interactions between them in order to reach this level, while Rule-Based Learners must make use of readable variables and obtain a number of rules that is manageable for humans.

Our aim is to propose a new extension of the well-known linguistic Fuzzy Rule-Based Systems (FRBS) [19–21] that is *specifically designed for regression problems*. This is done by combining new linguistic fuzzy grammar and a novel interpretable linear extension in order to obtain the highest possible transparency level, as this is our main objective, as well as a rule set of no more than 6 or 7 rules [22] for regression problems. A linguistic FRBS [20] makes use of fuzzy rules composed of linguistic variables [23] that take values in a term set with a real-world meaning, i.e., a variable whose values are words

drawn from a natural language that represent the basis for the concept of linguistic if–then rules.

On the other hand, the evolutionary fuzzy system (a fuzzy system designed by evolutionary algorithms [24]) is one of the greatest advances in the area of Soft Computing and subsequently, Computational Intelligence. Thus, the application of evolutionary algorithms for learning [25] the previously mentioned complex rule structures has been identified as being very useful in the context of XAI, since "machine learning methods based on evolutionary fuzzy systems preserve the original essence of comprehensibility exposed by Zadeh, also boosting their modeling abilities" [11]. In this contribution, the application of multiobjective evolutionary algorithms [26, 27] to learn understandable regression linguistic models, i.e., linguistic evolutionary fuzzy systems on regression problems, becomes a central axis where other techniques are also combined in order to enhance the learning process. However, even though linguistic models present great potential for transparency, and in particular simulatability, they still need to sacrifice part of their potential as too many rules are usually needed to obtain accurate predictions in real regression problems involving complex interpolated surfaces (continuous output variability in contrast to classification), large amounts of data, and/or complex data dependencies.

In this work, we not only propose a new extension of the fuzzy linguistic grammar, but also a novel interpretable linear extension which is specifically proposed for regression problems, together with an enhanced new linguistic tree-based evolutionary multiobjective learning approach in order to be able to describe a larger amount of regression data in terms that are as close as possible to those used by humans. The main aim is to allow the general behavior of the regression data covered, as well as their specific variability, to be expressed in a single rule. Thus, generally fewer rules would be needed to reasonably learn accurate regression linguistic models. Moreover, by being able to better summarize the extracted information, we can expect it to be richer and therefore more useful for human experts trying to find some insights in data whose real nature implies continuous variability. Even though there are a few new proposals for the design of interpretable models for classification problems, they are not directly applicable to those regression problems where it is quite difficult to model continuous complex surfaces with only a few rules that aim to separate the different values of the output variable. This is why our proposal actually aims to find and separate "tendencies," since they grasp the continuous nature in regression problems better. To our knowledge, there are no recent proposals that obtain really simple and transparent linguistic FRBSs (with only a small number of rules) for regression. More specifically, we aim to obtain linguistic regression models with approximately 5 rules without problematic overlapping (so that semantics are also preserved) and with competitive accuracy (so that reliability is maintained to high levels).

In particular, a new proposal extending the basic linguistic grammar usually used for predictive modeling (regression in our case) is presented in this paper. Similarly to the modified "OR" connector proposed in [28], it is based on the composition of new, more general linguistic term sets from single linguistic sets (strong fuzzy partitions or expert defined), but extends or modifies its syntactic representation in order to better resemble the way that humans might explain something (and does not repeat "OR" ... "OR" ... "OR" ... and so on, when you express a wider term). In this paper, this is referred to as a *Composed Fuzzy Linguistic Term Set* (CFLTS). Additionally, and this is of great importance for regression as it is key to maintaining competitive (or even improved) accuracy, a novel interpretable linear extension of the consequent rule structure (and the specific process to learn it) is proposed here, paying particular attention to interpretability criteria in order to explain the specific continuous variability of a rule that is possibly too general. It involves a new way of learning two parameters to help linguistically explain simple linear variations of the general behavior described by each rule, but only when they involve relevant accuracy improvements.

Moreover, we propose an enhanced multiobjective evolutionary algorithm (MOEA) in two stages (learning linguistic partitions and rules, plus tuning and rule selection) to optimize accuracy together with some well-known interpretability measures from the specialized literature that account for the number of rules and the overlap in the linguistic terms and/or rule inconsistency (the GM3M and RMI indexes [19, 21]). The main contribution in this algorithmic part is the inclusion of a new linguistic tree-based Rule Base (RB) learning algorithm that adapts perfectly to the new type of rules and therefore enhances the said evolutionary learning method, as node conditions can be fuzzified and the linear extension can be obtained at the tree leaves.

We have statistically tested the proposal in 23 regression datasets with different complexities (from 2 to 60 variables). The results obtained show the proposed method's effectiveness (obtaining less than 5 rules on average and no more than 7 for any of the 23 datasets) by applying Friedman's, Holm's and Wilcoxon's tests [29, 30] on all the interpretability indexes, as well as the number of rules and the accuracy, in order to compare them to some of the previous state-of the-art methods used to obtain interpretable pure linguistic FRBSs. Additionally, even though our proposal was mainly designed with interpretability purposes, it still presents highly competitive accuracy and it also compares to a state-of-the-art pure linguistic

accuracy-oriented method. Finally, we also analyze a case study in a real problem related to childhood obesity, where the analysis given of the obtained model is performed by a real expert.

Finally, a web page associated with this paper (https://www.ugr.es/∼ralcala/papers/ijfs21) has been developed that contains complementary material. It includes the following: the datasets collected and used in this study (the 5-fold cross-validation partitions); a brief description of the semantic interpretability indexes used (GM3M and RMI); an analysis of some representative examples of the linguistic models obtained in two of the benchmark problems; and, even though they are not comparable and are simply for benchmarking purposes, a comparison, from an accuracy point of view, between some representative highly accurate state-of-the-art general purpose models (Random Forest, etc.).

This contribution is organized as follows. The interpretability metrics that are used in this paper are set out in Sect. 2. Section 3 proposes the extension of the linguistic rule structure (new grammar and linguistic consequent extension). In Sect. 4, we present an effective MOEA to learn comprehensible linguistic FRBSs for regression problems together with a new linguistic tree-based learning to effectively learn the new type of rules proposed. Section 5 shows the experimental study on the proposed method, including statistical comparisons with some of the state-of-the-art interpretability and accuracy-oriented pure linguistic methods. It also includes a case study of a real problem related to obesity in children, where real experts analyze the obtained rules. Finally, Sect. 6 draws some conclusions.

2 Preliminaries: Interpretability Measures Considered to Ensure the Model Transparency

In [20], Gacto et al. define interpretability as "the capacity to express the behavior of the real system in an understandable way (comprehensibility)." According to this definition, the authors determine four different aspects that should be considered, and therefore measured, in order to obtain simple and interpretable linguistic fuzzy systems: Complexity and Semantics, at a linguistic label level or Data Base (DB) level, and at a RB level. This also resembles the definition in [31], where Magdalena defines interpretability by stating that "A fuzzy system is said to be interpretable if its reduced complexity and clear semantics make it possible for us to understand and explain its behavior by reading it." In order to take the four previously mentioned aspects into account, the following interpretability measures will be used and optimized in our proposal:

- Number of rules (Complexity at the RB level).
- Number of variables (Complexity at the DB level).
- GM3M [19, 21] (Semantic Interpretability at the DB Level). From 0.0 to 1.0 and with values close to 1.0 represent a really high proximity to the equally distributed strong linguistic partition.
- RMI [21] (Semantic Interpretability at the RB Level). From 0.0 to 1.0 and with values close to 1.0 represent the absence of inconsistent rules (i.e., high reliability level).

Even though GM3M and RMI are well-known and public descriptions are available (see [19, 21]), a brief description can be found in the web page associated with this paper (https://www.ugr.es/∼ralcala/papers/ijfs21). In any case, we only introduce them and their aim (what are they measuring and why) so that their particular formulations can be found in the corresponding papers in which they were proposed.

3 New Grammar Proposal: Extending the Classic Linguistic Rule Structure

In this paper, we propose a new extension of the classic linguistic rule structure in order to allow more general representations of certain parts of the data, as well as a more specific explanation and follow-up for human beings. The main aim is to simplify the regression model by explaining the general behavior of the data and, at the same time, explaining the variability of the data within the scope of the general description provided.

Therefore, we can focus on understanding the different general explanations and then see what happens, particularly in the general continuous domain that the rule is explaining. This way of describing the reality of continuous numerical data is not very far from how humans explain themselves, and in fact we rely on it to introduce a dialogue to illustrate explanatory concepts throughout the article. When a person explains something related to continuous numerical values, they first tend to generalize, and then explain, if needed, what happens more specifically in the context of the fact that they have just mentioned. Furthermore, in many cases, this is usually relativized with respect to the explanation given initially.

For instance: If a person's height is from "high onwards," then the amount of calories consumed will be between high and very high (which is around 2600 kcal), but... the amount of calories consumed should be reduced by 7.3 kcal per year when subjects are over 25 years of age. This does not only explain the differences between the particular cases and the general rule, but also provides us with additional and useful information since age is

identified as an important factor related to calorie consumption, and 7.3 would be its specific relationship to people that have a "high onwards" height.

If we are able to come up with the formulation that corresponds to this way of explaining the continuous numerical reality, we should be able to obtain much more explanatory and simpler models with a good level of precision, and thus make a leap toward Simplicity + Semantic Interpretability + Precision, instead of the classical Simplicity vs. Semantic Interpretability vs. Precision. In fact, the fundamental trade-off between accuracy and interpretability is identified by some authors as one of the possible myths in machine learning [14], even though improving both of them at the same time is actually a great challenge and makes the trade-off almost impossible.

3.1 New Grammar to Learn More General Linguistic Rules

Here, we propose how to combine and interpret new broader linguistic expressions in the context of predictive modeling, namely CFLTSs (composed fuzzy linguistic term sets), and the operators required for modeling. Similar to the modified "OR" connector proposed in [28], CFLTS is a syntactic extension based on the composition of new and more general linguistic term sets from single linguistic sets (strong fuzzy partitions or expert defined), however, it also modifies the "... OR ... OR ... OR ..." syntactic representation in order to help it resemble the way that humans explain. Thus, we can consider the proposal in [28] to be an initial version of CFLTSs.

As such, and as we are trying to grasp and linguistically explain the different and more general situations that should be described in each particular data region, more general or broader statements would be achieved. We therefore define some new specific CFLTSs in an attempt to provide the consistency of the inference system and complete coverage of those aforementioned regions. Of course, we will also take into account the fact that the new grammar defined for the fuzzy rules should still resemble the natural way in which people speak. For example: We should be able to clearly indicate that the linguistic term of one of the variables represents values from the average onwards, e.g., a person has a height that is from medium onwards (i.e., it is medium, high, or very high).

The basic form of a fuzzy rule is defined as follows: IF *antecedent* THEN *consequent*. This type of rule must work correctly with both equally distributed and free or tuned membership functions (MFs). See an example of both types in Fig. 1 by considering triangular MFs, where l and r define the corresponding domain $[l, r]$ of the variable values.

In the case of linguistic FRBSs, the antecedent is a linguistic fuzzy proposition composed of atomic linguistic

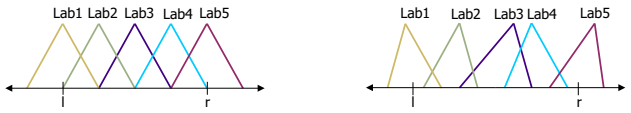

Fig. 1 Initial labels (equally distributed *vs* free/tuned MFs)

fuzzy propositions (or linguistic term sets) (AtLingFuzzyProp) which are connected by the AND connective (using the minimum as a t-norm operator). It has the following form: *AtLingFuzzyProp* AND *AtLingFuzzyProp* ... AND *AtLingFuzzyProp*. Each atomic fuzzy proposition is associated with a single linguistic variable. We introduce or propose the (classic or CFLTSs-based) linguistic propositions used by showing some graphical examples of the example linguistic partitions from Fig. 1:

- **X_i is Lab3** in Fig. 2 (classic linguistic proposition). The membership degree of values in X_i to the linguistic term Lab3 determines the degree to which this atomic/single fuzzy proposition is verified.
- **X_i is *up to* Lab2** in Fig. 3 (proposed new CFLTSs). It is a composed term set or proposition which represents Lab2 and all the smaller values. The main difference between this set and the previous one is that it has a membership degree that is equal to 1.0 from $-\infty$ to the center of the label Lab2, where the center of the label is the central point of the label core (i.e., the core itself when we have triangular MFs).
- **X_i is *from* Lab3** in Fig. 4 (proposed new CFLTSs). It is a composed term set or proposition that is similar to "up to" but representing Lab3 and all the greater values. It has a membership degree equal to 1.0 from the center of label Lab3 to $+\infty$.
- X_i **is *from*** Lab2 ***to*** Lab4 in Fig. 5 (proposed new CFLTSs). It represents Lab2, Lab4, and all their intermediate values. It has a membership degree equal to 1.0 from the center of label Lab2 to the center of Lab4. It not only allows sets above or below a given point, but also allows those in a linguistic intermediate range. Example: If the temperature is from low to medium, then do not activate the air conditioning (considering a linguistic partition with five labels starting in "very low" and ending in "very high").

Furthermore, considering that we use the First Infer Then Aggregate (FITA) inference system, we also allow the consequent proposition of a rule to be expressed as the intermediate output between two adjacent labels ("between" was proposed in [32]) as it is perfectly

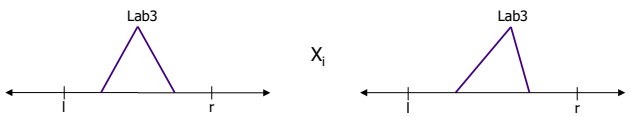

Fig. 2 Associated MF for the single linguistic term Lab3

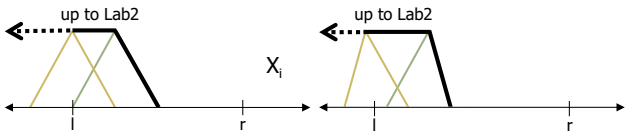

Fig. 3 Composed MF for the *CFLTS* "up to" Lab2 for input variables (X_i)

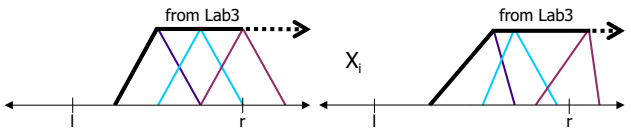

Fig. 4 Composed MF for the *CFLTS* "from" Lab3 for input variables (X_i)

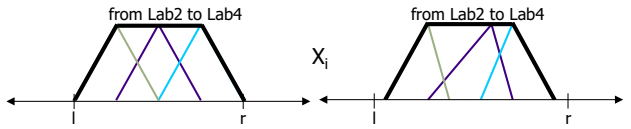

Fig. 5 Composed MF for the *CFLTS* "from" Lab2 "to" Lab4 for input variables (X_i)

comprehensible and commonly used by human beings. It does so, however, in a different manner to the previous CFLTSs. **Y is** *betw.* **Lab2** *and* **Lab3** directly represents the mid-point between these two labels, which makes more sense and allows for more flexible and accurate rules. Figure 6 depicts this linguistic composition, which, as it only makes sense for the rule consequent, is applied exclusively to the Y output variable labels or term set.

Finally, we provide two examples of the final rule structure, demonstrating the use of all these new types of propositions:

$$\begin{array}{r} \textit{IF } X_2 \textit{ is up to Lab3 AND } X_5 \textit{ is from Lab2 to Lab4} \\ \textit{THEN } Y \textit{ is betw. Lab1 and Lab2} \\ \textit{IF } X_2 \textit{ is from Lab4 AND } X_5 \textit{ is Lab3} \\ \textit{THEN } Y \textit{ is Lab3} \end{array} .$$

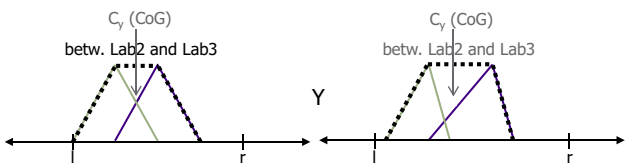

Fig. 6 Intermediate output Center of Gravity (CoG) from the composed MF for the *CFLTS* "betw." Lab2 "and" Lab3 for the output variables (Y)

3.2 Rule Extension Proposal for Explaining Specificity in Regression

Thanks to the new grammar, it is possible to learn more general rules that fit perfectly with the generality of the data. However, this would lead to very low accuracy models that only represent generalities. The greatest difficulty is to find a way for a rule that explains a more general concept to also explain the relative specificity of each of the data it describes.

In this section, we propose an extension of the linguistic rule structure that can be also linguistically interpreted. It is based on the ideas expressed at the beginning of Sect. 3 (see the calories example). In order to implement this new way of expressing more specific concepts that are based on (or relative to) more general ones, a rule with the following structure is proposed:

$$\begin{array}{r} \textit{IF } X_i \textit{ is from Med} \\ \textit{THEN } Y \textit{ is betw. Low and Med} \\ \textit{with } \pm s \textit{ per } \pm \textit{unit in } X_r \textit{ from } d \end{array} ,$$

where "from Med" expresses that the X_i variable takes values above Med and "betw. Low and Med" indicates that variable Y takes values between the two adjacent labels Low and Medium. The extended part of the rule is based on obtaining a line whose abscissa values are centered on the output of the general part of the rule, so that s is the amount that we must add/subtract to the general output of the rule (C_y) per unit above/below d at the X_r value of the input example for which we are estimating the output.

Let us remember the example given at the beginning of Sect. 3: If a person's height is from high onwards, then the amount of calories consumed will be between high and very high (which is around 2600 kcal), but... the amount of calories consumed should be reduced by 7.3 kcal per year over the age of 25 years old, or be increased by 7.3 kcal per year under the age of 25. It then could be expressed as:

$$\begin{array}{r} \textit{IF Height is from High} \\ \textit{THEN } Y \textit{ is betw. High and VeryHigh} \\ \textit{with } \mp 7.3 \textit{kcal per } \pm \textit{year in Age from } 25 \end{array} .$$

This rule is a formal representation of the explanations given in the example, which is based on the use of the proposed CFLTSs and the computation of the corresponding straight line. However, it can still be read by using the same linguistic expressions from the example.

As explained in the previous subsection, once the general part of a rule that provides a general rule output (C_y) is determined, the particular values of s (slope or coefficient of X_r) and d (displacement in X_r that forces the intercept to

coincide with the central point of the general rule consequent label, namely c_y)[2] are obtained based on the least squares estimation of the straight line in the examples (X_r, Y). Let us consider the formula of the so obtained line:

$$Y' = \beta_0 + \beta_1 X_r,$$

where β_0 and β_1 are the two unknown constants of the line obtained by least squares. To obtain s and d, so that the explained interpretation is correct ($Y' = s \cdot (X_r - d) + c_y$), we must equate both equations. Taking into account that both lines must have the same slope, $s = \beta_1$, the following equalities can be established:

$$\beta_1 \cdot (X_r - d) + c_y = \beta_0 + \beta_1 X_r.$$

So that the same slope can be set and this equality can be solved, s and d can be obtained as:

$$s = \beta_1,$$
$$d = (c_y - \beta_0)/\beta_1.$$

When X_r is also used in the rule antecedents, d must be within the examples that the rule antecedent covers for that variable, since it must explain the variability within its range.

4 Proposed Learning: A New Tree-Based Hybrid Evolutionary Multiobjective Algorithm

In this section, we propose a new learning methodology that adapts to the new type of rule structure explained in Sect. 3.1. Our aim is to design an algorithm that uses the new rule representation to obtain linguistic FRBSs as comprehensively as possible, while also maintaining or improving their accuracy (reliability). Likewise, it should be able to minimize the number of rules, maximize the linguistic interpretability (both GM3M and RMI), minimize the error, and keep the length of the rules (number of conditions) within a reasonable limit.

It is based on a two-stage tree-based hybrid evolutionary multiobjective algorithm:

- First stage: Learning of the initial linguistic partitions and the associated linguistic rules. This stage performs an embedded multiobjective evolutionary DB learning, minimizing both the number of rules and errors. This is a multiobjective evolutionary process that learns the DB and wraps a fast method to derive a set of rules for each DB definition. In this paper, we have put forward hybridization with a new linguistic tree-based rule learning in order to profit from the newly proposed rule structure. This is done by extending the well-known M5 -prime [33, 34] as the method for deriving a set of rules for each of the evolved DBs.
- Second stage: Post-processing stage to further refine the learned solutions. This is a multiobjective evolutionary algorithm that fine-tunes the MFs and rule selection, which helps minimize the number of rules and maximize the linguistic interpretability of both GM3M and RMI, and minimize the error of the simple global structure obtained in the first stage (initially based on strong fuzzy partitions).

First, we propose the new linguistic tree-based rule learning method to account for generality by briefly describing the M5 -prime algorithm and how we have adapted it to learn linguistic rules based on the use of CFLTS and the consequent extension, which is proposed to account for specificity. Finally, the main characteristics of both stages of the proposed method are presented.

4.1 Proposed Linguistic Tree-Based Rule Learning

In this section, we propose a new linguistic tree-based algorithm for learning the corresponding extended rule set, considering the previously proposed extended grammar, and based on the existence of a well-defined linguistic partition. The proposed algorithm is based on the way the well-known M5 -prime [33, 34] algorithm works.

In the following, we first briefly describe the M5 -prime algorithm [33, 34], and then explain the modifications that have been made.

4.1.1 Preliminaries: Brief Description of the M5-Prime Algorithm

The M5-prime [33, 34] method is a regression tree (specifically a *model tree*), which means that it is used to learn a tree whose node leaves include local multivariate linear models in order to predict the values of a numerical response variable Y. That is, while the M5-prime tree uses the same approach as the well-known CART tree to choose the mean square error as a function of impurity, it does not assign a constant to the leaf node but instead adjusts to a multivariate linear regression model; the model tree is, therefore, analogous to multivariate linear functions in parts. The M5-prime tree can learn efficiently and can handle very high dimensional problems—up to hundreds of attributes, making it a fairly fast method. This capacity differentiates M5-prime from other regression trees such as MARS, whose costs grow very quickly when the number of characteristics increases. In addition, the advantage of M5-prime over CART is that the models are generally much smaller than those obtained by the regression trees and tend to be more accurate.

[2] As mentioned, we are considering a FITA inference based approach and Center of Gravity (CoG).

Generation of the M5-Prime Tree.

The M5-prime method follows the same recursive node division strategy as decision trees. Suppose that we have a collection T of training examples. The set T is associated with a leaf, or a split is chosen based on same test that divides T into subsets corresponding to the best division on the said test. Then, the same process is applied recursively to the subsets. This process often produces overlearned structures that must be subsequently pruned.

The information gain in the M5-prime tree is measured by the reduction of the standard deviation before and after the division. The first step is to calculate the standard deviation of the output variable values of the example data in T. Unless T contains very few cases or its values vary only slightly, T will be divided according to the test results for each possible cut point (for every value at each attribute). Let T_i denote a subset of examples corresponding to the i-th result according to a specific split. If the standard deviation $sd(T_i)$ of the output variable values of example data in T_i is treated as an error measure, the expected reduction of the error can be written as follows:

$$\triangle\text{error} = \text{sd}(\text{T}) - \sum_{\text{i}} \frac{|\text{T}_\text{i}|}{|\text{T}|}\text{sd}(\text{T}_\text{i}).$$

Then the M5-prime tree will choose the split that maximizes the expected error reduction. For comparison, CART chooses a division to give the largest expected reduction, either variance or absolute deviation.

M5-Prime Pruning.

Pruning is carried out from the leaves to the root node. At each internal node, the M5-prime tree compares the estimated error of that node and the estimated error of the subtree below it. Thus, the subtree is pruned when it does not improve the said node's performance.

The model tree M5-prime uses a pruning method based on estimated errors. The key factor of this method is how it estimates the model error in unseen input data, since overfitting directly depends on it. The M5-prime tree calculates it by first averaging the absolute difference between the output values of the training data and their predicted values. This will generally underestimate the error in unseen data, so M5 -prime multiplies it by $(n + v)/(n - v)$ where n is the number of training cases and v is the number of parameters in the model [33, 34]. The effect aims to increase the estimated error of models with many parameters obtained from a small number of examples. More efficiently, the estimated error of a node model is calculated as the sum of the estimated error of the left and the right subtrees, multiplied by the proportion of samples descending to each of them, respectively.

Linear Models.

A multivariate linear model fits to the training data in each node of the tree using standard regression techniques. After the full size tree is produced, a multivariate linear regression model is fitted for each node on its associated training examples by following a backward operation mode. It starts with the leaf nodes, considering only the variables used in the splits from the root to the leaf node. However, in the case of an internal node, it is restricted to those variables that are referenced by the splits or the linear models in the subtrees below the node. M5-prime compares the error estimates of a linear model with those of the pruning subtrees, so it allows for fair competition conditions where these models also use the same information.

After learning a linear model, M5-prime simplifies it by greedily eliminating coefficients to minimize its estimated error one by one. In general, this could result in an increase in the average residue; however, it also reduces the previous multiplicative factors, so the estimated error may decrease.

4.1.2 Proposal: Extension of M5-Prime to Generate Linguistic Trees

In this article, we propose an adaptation of the M5-prime-Rules method which generates a set of rules from the M5-prime tree. Let us assume that we have a previously defined DB, i.e., the linguistic partitions for all the variables involved in the problem being solved, where l_j^v is the j-th linguistic term of the v-th variable.

Proposed Modifications To generate linguistic trees based on the extended grammar and rule structure proposed, the following changes are made to M5-prime:

1. Changes in the tree construction:
   - Calculate the standard deviation, $sd(T)$ or $sd(T_i)$[3], weighted by the pairing of the examples with the rule built up to the node in question. That is, the corresponding proposition for the path from the root to the T_i split is built, and the pairing of each example in T_i to this proposition is calculated and considered as its corresponding weight. The standard deviation $sd(T_i)$ of the target/output values of the example data in T_i is computed as the square root of the weighted average of the differences of each output value from its mean.
   - Determine all the possible splitting points, j, so that the left branch would be associated with "up to l_j^v" and the right one with "from l_{j+1}^v" for all the input v variables from the existent linguistic partitions. Take into account that, when v has been previously

[3] T_1 and T_2, left and right, since we only have binary splits.

used at any split in the path from the root to the node, we should only try all the possible splits in the form "from l^v_{ini} to l^v_j" (left branch) and "from l^v_{j+1} to l^v_{end}" (right branch), where l^v_{ini} and l^v_{end} are extreme linguistic terms determined by the previous divisions of v on the said path.
- For each possible split, move the examples in T to the corresponding T_i subsets by considering their pairing with the composed statements/propositions on the left and right[3]. Calculate their respective standard deviations $sd(T_i)$ weighted by the pairing of the examples with the corresponding proposition.
- Choose the split by maximizing the expected error reduction, $\triangle$error.
- Obtain the tentative specific parts (Sect. 3.2) of the variable that best fits the data of the rule associated with the node and the child nodes, and the correlation coefficients, R^2, for the straight lines they represent.
- A branch node is considered to be a leaf if it has less than 4 examples, or if the standard deviation of the branch is less than a small fraction of the total (ie, $sd(T_i) <= 0.05 * sd(CompleteDataSet)$), or if the R^2 of the line represented in the parent's consequent is greater than that of the branch node.
- Additionally, the depth of the tree is limited to three, since our main objective is to obtain the simplest possible systems with the least number of rules and minimum rule lengths.
- The algorithm would continue splitting this way until the entire tree has been obtained.

2. Changes in the tree pruning. Pruning is performed as it was in the original M5-prime for each non-leaf node, starting near the bottom. However, in this case, it is based on computing the Mean Square Errors (MSEs) of a given node and its subtree. If the error of a node is less than the error of its subtree plus a small percentage (error becomes 5% worse), this subtree is pruned and the node is set as a leaf. Thanks to this additional percentage, the tree avoids the excessive adjustment to the data. It continues until the pruning has been completed.
3. Additionally, convert the tree into rules. Build a rule for each possible path from the root to the leaves.

Specificity Component of the General Rules Obtained.

For each rule learned, the algorithm can generate (or not) its specific part about one of the variables available in the dataset. To obtain the parameters involved in this part of the rule, the procedure described in Sect. 3.2 is applied to each of the input variables of the training dataset. The chosen variable and the computed associated parameters are those with the best R^2 for the straight line they represent. This part, however, is not added to the rule if the best R^2 is below 0.25.

In addition, it has been taken into account that, when the chosen variable is also a part of the rule antecedent, the obtained parameter d must be within the range of the examples covered by that rule in the said variable. Therefore, the specific part of the rule will not be added when d is out of this range. In this case, the associated R^2 would be set to zero.

Summarizing everything explained in this subsection, Fig. 7 shows a flowchart of the linguistic tree-based rule learning method in order to make its steps more understandable.

4.2 Hybrid Multiobjective Evolutionary Learning

In this section, we present the proposed two-stage hybrid multiobjective evolutionary algorithm. It is partially based on the (FS-MOGFS) algorithm from [35], which applies a modified version of the SPEA2 selection scheme [36] including the corresponding external population. The basic execution scheme of the proposed methodology (including the objectives of each stage) is depicted in Fig. 8. Since the whole algorithm is specifically designed to search for the highest possible interpretability while always focusing on the most accurate models, the most accurate solution is considered to be the final output from both stages. Please also take into account that Fig. 7 in the previous section shows a flowchart that clarifies the operation of the process for new linguistic tree-based rule learning.

As previously mentioned, this algorithm is made up of two stages, which are described in the following.

4.2.1 First Stage: Learning the Knowledge Base

This stage performs an embedded multiobjective evolutionary DB learning, which wraps the proposed linguistic tree-based rule learning method. The components needed to implement this stage are explained in depth below. They are: coding, rule base derivation, objectives, population initialization, crossing and mutation, incest prevention, restarting, and stopping condition.

Coding of the DB.

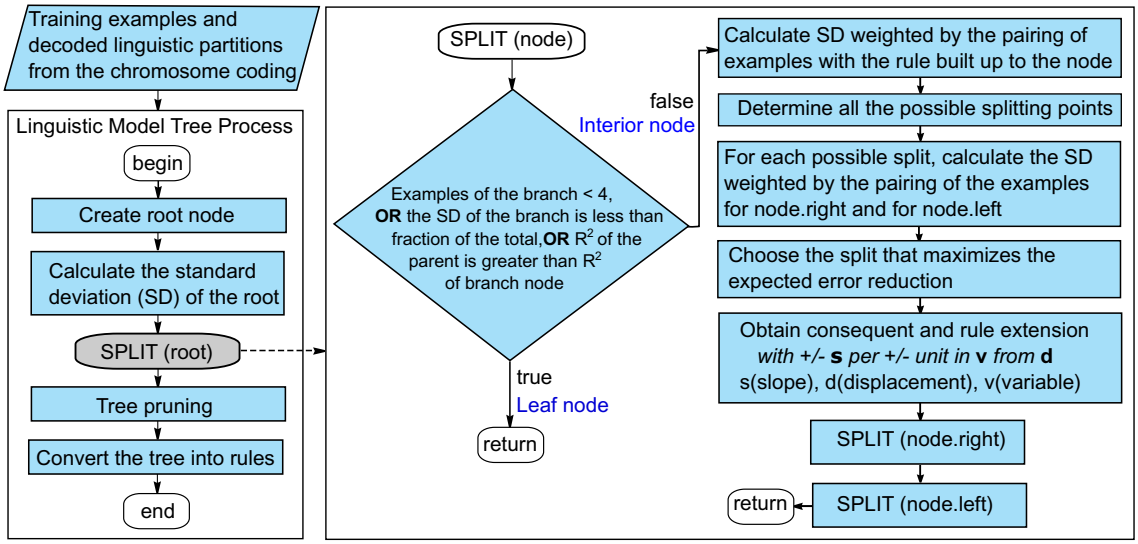

Fig. 7 Flowchart of linguistic tree-based rule learning

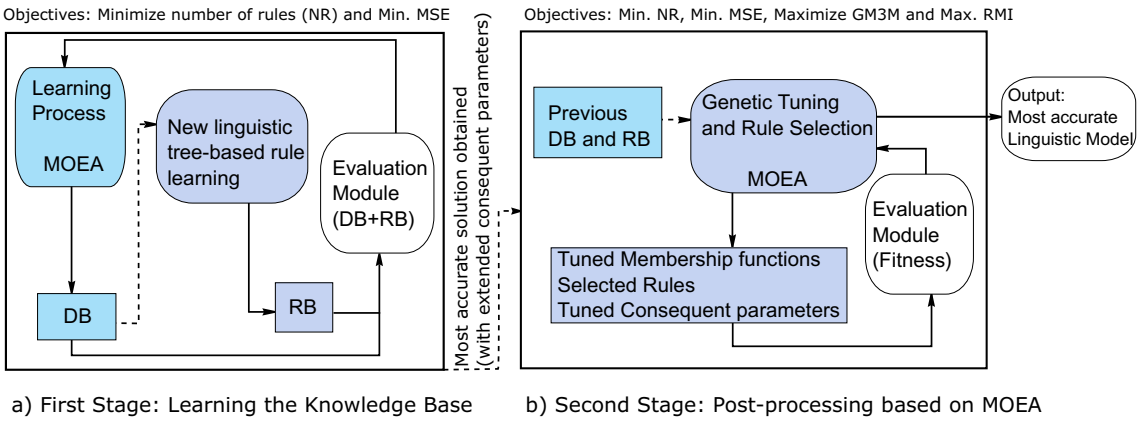

Fig. 8 Flowchart of the proposed method

A double coding scheme ($C = C_1 + C_2$) is applied to represent both parts, *granularity* (number of simple linguistic terms) and *tuning of parameters* (considering the classic triangular-type MFs):

– Number of labels (C_1): This part is a vector of integers with size N (with N representing the number of linguistic variables) in which the granularities of the different variables are encoded,

$$C_1 = (L^1, \ldots, L^N).$$

Each gene L^i represents the number of labels used by the variable i-th and takes values in $\{2, \ldots, 7\}$. In [22], cognitive psychologist George A. Miller proposed the "magic number," stating that the number of different objects (in our case labels) that can be handled in short-term memory is 7 ± 2, which is the number of conceptual entities a human being can handle at one time. This number of maximum labels has been historically assumed in the linguistic FRBSs area. As for the input variables, they can also take a value equal to 1 to determine that the corresponding variable is not used.

– Lateral displacements (C_2): This part is a vector of real numbers with size N in which the displacements of the different variables are coded (see an explanation of the lateral adjustment of fuzzy partitions in [35]). Thus, C_2 contains the displacement values, where each gene is the particular displacement value of the fuzzy partition of the corresponding linguistic variable. It takes values in $[-0.1, 0.1]$,

$$C_2 = (\alpha^1, \ldots, \alpha^N).$$

Rule Base Derivation. To obtain a complete linguistic model from a given chromosome (i.e., an evolved DB or whole linguistic partition), we apply the proposed linguistic tree-based rule learning method (see Sect. 4.1.2) to the DB encoded by this chromosome.

To decode this DB before learning the rules, strong fuzzy equally distributed partitions are defined with granularity values in C_1. Secondly, the MFs of each variable move slightly and uniformly to their new positions over the displacement values in C_2. The linguistic tree-based rule learning algorithm is applied to this DB in order to obtain its associated RB (see the type of rule structure and an example in Sects. 3.1 and 3.2).

Objectives.

Once a complete linguistic model is obtained, the following two objectives are minimized: the number of rules (simplicity) and the MSE (precision),

$$\text{MSE} = \frac{1}{2 \cdot |E|} \sum_{l=1}^{|E|} (F(x^l) - y^l)^2,$$

where $|E|$ is the dataset size, $F(x^l)$ is the output from the FRBS obtained from a given chromosome when considering the example l-th, and y^l is the known desired output. The fuzzy inference system considered to obtain $F(x^l)$ is the center of gravity weighted by the matching strategy as a defuzzification operator and the minimum t-norm as implication and conjunctive operators.

Initialize the Gene Population.

The initial population is composed of two different subsets of individuals in order to ensure the inclusion of the possible combinations between antecedent and consequent granularity, while also including random maximum diversity:

- In the first subset, each chromosome has the same number of labels for all input variables in the system. To provide diversity in C_1, these solutions have been generated considering all the possible combinations of the antecedent part, that is, from 2 labels to 7 labels in all the input variables (6 combinations). For each of these combinations, all possible combinations are generated in the corresponding consequent part (6 combinations for each input combination). In addition, for each of the above combinations, two copies are included with different values in the C_2 part. The first is included with random values in $[-0.1.0.0]$ and the second with random values in [0.0, 0.1] . Therefore, a total of 72 $(6 * 6 * 2)$ different individuals are generated. If there is no space for these solutions, they will range from the smallest granularities (the most interesting combinations) to the highest possible granularities.
- In the second subset, we generate random solutions in order to complete the population (values in $\{2, \ldots, 7\}$ for C_1 and values in $[-0.1, 0.1]$ for C_2).

Finally, except for problems with less than 3 input variables, an input variable v is eliminated at random, $L^v = 1$, in the first individual. This action is repeated until there are no more than 5 variables in this individual. If the problem does not have more than 5 variables, this action is not repeated, so that only one variable is eliminated at random. This process is applied to all individuals in the population to avoid the generation of solutions with an excessive number of rules.

Crossover and Mutation Operators.

The crossover operator depends on the part of the chromosome to which it is applied. A crossing point is randomly generated and the classic crossover operator is applied to C_1. The Parent Centric BLX (PCBLX) operator, which is based on BLX-α, is applied to C_2. PCBLX is described more specifically below. Suppose that $X = (x_1 \ldots x_n)$ and $Y = (y_1 \ldots y_n)$, with $x_i, y_i \in [a_i, b_i] \subset \Re$ y $i = 1 \ldots n$, are two chromosomes with real coding that are going to be crossed. PCBLX generates the following two children:

- $O_1 = (o_{11} \ldots o_{1n})$, where o_{1i} is generated randomly (uniformly) in the range $[l_i^1, u_i^1]$, with $l_i^1 = \max\{\mathrm{a_i}, \mathrm{x_i} - \mathrm{I_i}\}$, $u_i^1 = min\{b_i, x_i + I_i\}$, and $I_i = \mid x_i - y_i \mid \cdot \alpha$. In our case, α has been set to 0.3.
- $O_2 = (o_{21} \ldots o_{2n})$, where o_{2i} is generated randomly (uniformly) in the range $[l_i^2, u_i^2]$, with $l_i^2 = \max\{\mathrm{a_i}, \mathrm{y_i} - \mathrm{I_i}\}$ and $u_i^2 = min\{b_i, y_i + I_i\}$.

As such, four new individuals are obtained by combining the two children generated from C_1 with the two children from C_2. The mutation operator is applied to each of them with probability P_m. The mutation operator decreases the granularity by 1 in a randomly selected gene g $(L^g = L^g - 1)$, or randomly determines a greater granularity in $\{L^g + 1, \ldots, 7\}$ with the same probability. No decrease occurs when causing DBs with a single input variable. The same gene is also changed randomly at C_2. Finally, after mutation, only the two most accurate individuals are used as descendants.

Incest Prevention.

An incest prevention mechanism has been included following the concepts of CHC [37] and taking only C_2 into account. Following the original CHC scheme (for binary coding), two parents are crossed if their hamming distance divided by 2 is above a predetermined threshold, L. Since C_2 makes use of a real coding scheme, we have to transform each gene by considering a Gray Coding (binary code) with a fixed number of bits per gene (*BITSGENE*), which is determined by the system expert. In this way, the threshold value is initialized as:

$$L = (\#GenesC_2 \cdot BITSGENE)/4.0.$$

Typically, L is decreased by one when there are no new individuals in the next generation. In our case, to accelerate

convergence L will be reduced by two in each generation. Incest prevention represents a way to provide a good balance between exploration and exploitation, avoiding unnecessary crossings of very similar solutions in the early stages of the algorithm.

Restarting and Stopping Condition.

In order to move away from local optima, a restarting mechanism is applied by emptying the external population, including the most precise individual in the new population and generating the remaining individuals at random (taking values between 1 and the granularity coded in the most accurate individual for each gene in C_1). This mechanism is applied when the threshold value L is below zero (L is set to its initial value).

The algorithm ends when a maximum number of evaluations is reached or when L is below zero for the second time. This means that only two exploration/exploitation stages are needed to achieve convergence.

4.2.2 *Second Stage: Post-Processing Based on Multiobjective Evolutionary Algorithms*

Once a candidate linguistic model has been generated in the learning stage, the next step is to apply a post-processing performance tuning of the database and the extended consequent parameters, and a rule selection. It is based in part on the post-processing algorithm presented in [35] and it once again applies a modified version of the SPEA2 selection scheme [36] including the corresponding external population. The new method is designed to adapt to the type of rule generated in the first part that allows some input variables, including CFLTSs, and thus making use of the composed terms: "up to," "from," "from-to," "betw.-and," and the parametric values of the extended rule consequent. The components needed to implement this stage of the algorithm are explained in depth below.

Objectives.

Since the main objective is to improve the interpretability of linguistic FRBSs as well as their accuracy (reliability), there are several metrics that will be used to evaluate each FRBS. Again, the MSE and the number of rules are used for both precision and complexity. Semantic interpretability will be evaluated using one of the two previously presented metrics, GM3M or RMI (see Sect. 2).

Therefore, each population chromosome, representing an FRBS, will be evaluated according to the degree of compliance in each of the following four objectives:

1. **Maximize** the value of the **GM3M** index: To preserve the semantic interpretability of the initial MFs.
2. **Maximize** the value of the **RMI** index: To preserve or improve the semantic interpretability of the rules.
3. **Minimize** the number of rules **NR**: To reduce the complexity of the model.
4. **Minimize** the mean square error **MSE**: To reduce the error of the system (improving its reliability).

Coding Scheme A triple coding scheme will be used, one for the *rule selection* (C_S), one for the tuning of MFs (C_T) and another one for tuning of the extended consequent linear parameters (s and d) (C_L). A chromosome p has the form $C^p = C_S^p + C_T^p + C_L^p$.

The coding for $C_S^p = (c_{S1}, \ldots, c_{Sm})$ consists of a binary vector with size m (number of initial rules). Depending on whether a rule is selected or not, the corresponding gene takes the values '1' or '0,' respectively.

For C_T, we use a vector of real numbers that represents the characteristic values of the MFs. That is, all the simple MFs are adjusted, which also involves an indirect adaptation of the CFLTSs used in the rules as they are based on the simple MFs. The Coding of $\mathbf{C_T}$ is:

$$\begin{aligned} &C_T^p = C_1 C_2 \ldots C_n; \\ &C_i = (a_1^i, b_1^i, c_1^i, \ldots, a_{m^i}^i, b_{m^i}^i, c_{m^i}^i), i = 1, \ldots, n. \end{aligned} \tag{1}$$

C_L is a vector of real numbers with size $m * 2$ that represents the extended consequent linear parameters (s and d) for each of the m rules. See this type of rule structure and an example in Sects. 3.1 and 3.2. $C_L^p = (c_{s_1}, \ldots, c_{s_m}), (c_{d_1}, \ldots, c_{d_m})$.

In the learning phase, the range of acceptable values for these parameters (which are used as gene domains) is obtained, as well as the linear parametric part associated with each rule consequent.. For example, the range of possible acceptable values for s (slope) is obtained without the line leaving the data range.

Initial Population

The initial population is obtained by following the rules detailed below:

- C_S: all the genes of all the individuals take value '1,' so that rule removal is carried out progressively at early stages, mainly via mutation.
- C_T of the first individual: the coding of the MFs of the initial model is directly introduced.
- C_T of the remaining individuals: it is initialized randomly, taking into account the classic ranges of variation established in the literature for triangular-shaped MFs (see [35]).
- C_L of the first individual: the parameters of the rule lines (s and d) of the initial model are directly introduced.
- C_L of the remaining individuals: it is initialized randomly, taking into account the range of variation of the extended consequent linear parameters previously calculated in the first stage.

Crossover and Mutation Operators.

The intelligent crossover operator and the mutation operator used in this proposal have been selected based on previous experiences [35] in dealing with the particular problem of rule selection and MF tuning. When using two different types of coding in a chromosome, it is necessary to define specific operators for each of the parts of the chromosome. The steps for obtaining each of the descendants are shown in the following:

- The C_T and C_L parts of the descendant are fixed by using the well-known BLX-0.5 [38] crossover operator.
- The C_S part of the descendant is calculated by applying the intelligent crossover operator to rule selection problems in [35] once the real part C_T of the descendant has been obtained.

This process is repeated until the four descendants are obtained. Once the descendants have been generated, the mutation operator is applied. This operator is applied independently to each of the chromosome parts. For the real parts, C_T and C_L, the operator changes the value of a randomly selected gene taking into account the ranges of variation of the parameters. On the other hand, in the C_S part, the operator sets another selected randomly gene directly to '0.' After applying the mutation operator, only the two descendants with the best precision are finally selected.

Two problems are solved by applying these operators as previously described. First, the result is more productive when individuals that contain different rule configurations are crossed. Second, rule extraction is favored from the moment the mutation only focuses on eliminating unnecessary rules.

Special Mechanisms to Handle Balance Precision-Interpretability.

The proposed algorithm uses the selection mechanism of *SPEA*2 [36]. As we have previously mentioned, the following modifications have been included to improve its search capabilities:

- An **incest prevention mechanism** based on the concepts of CHC [37]. The method uses the mechanism as described in [19]. However, in this case, we only prevent a premature convergence in the C_T part. It performs as follows: only those parents whose Hamming distance divided by 4 is greater than a threshold are crossed. As a real coding scheme is used (only the C_T part is considered), it is necessary to transform each gene to Gray code with a fixed number of bits per gene (*BGene*), as determined by the expert. In this way, the threshold value is initialized as $L = \frac{\#C_T * BGene}{4}$, where $\#C_T$ is the number of genes from the C_T part of the chromosome. In each generation of the algorithm, the value of the threshold decreases by one unit, progressively allowing closer solutions to be crossed with one another.
- The **restarting operator**, which forces the external population to be emptied, generates a new initial population that contains some of the best solutions already located by the algorithm. Specifically, the new external population contains a copy of the individual with the best accuracy and copies of the two individuals with the best value in each of the other objectives (GM3M, RMI and NR). In total, it keeps 7 of its individuals, while the rest of the individuals reinitialize, taking the same values as the individual with the best precision in C_S and random values in C_T and C_L. This reinitialization process is applied when the threshold L is below zero. In addition, this reinitialization process is deactivated in the final evaluations of the algorithm, and if it has never been applied before, then it is deactivated in half the total number of evaluations.

5 Experimental Study

In this section, we will evaluate the goodness of the proposed grammar and the rule structure extensions from an interpretability point of view, while also paying special attention to the accuracy (reliability). We also evaluate the usefulness of the proposed method presented in this document. This section is organized as follows:

1. Sect. 5.1 introduces the experimental setup.
2. Sect. 5.2 shows a statistical comparison with state-of-the-art linguistic methods, which also optimize semantic interpretability measures.
3. Sect. 5.3 presents a statistical comparison with a state-of-the-art pure linguistic accuracy-oriented method in order to show how the proposed method achieves a good level of accuracy while it also obtains really simple and semantically consistent models.
4. Sect. 5.4 presents a case study on a real problem related to childhood obesity where real experts interpret the obtained model.

5.1 Experimental Set-Up

In this subsection, we first show the experimental setup used in this paper. The experiments are carried out with several regression datasets. The main characteristics of these datasets are presented in Table 1: name of the dataset (NAME), short name or acronym of the dataset (ACRO), number of variables (VAR), and number of examples (CASES). The experimental study is carried out with 23

Table 1 Regression datasets

NAME	ACRO	VAR	CASES
Abalone	ABA	8	4177
Anacalt	ANA	7	4052
Baseball	BAS	16	337
Boston housing	BOS	13	506
Diabetes	DIA	2	43
Machine CPU	CPU	6	209
Electrical Maintenance	ELE	4	1056
Body fat	FAT	14	252
Forest Fires	FOR	12	517
Friedman	FRI	5	1200
Mortgage	MOR	15	1049
Auto Mpg 6	MPG6	5	392
Auto Mpg 8	MPG8	7	392
AutoPrice	PRI	15	159
Quake	QUA	3	2178
Stocks domain	STP	9	950
Strike	STR	6	625
Treasury	TRE	15	1049
Triazines	TRI	60	186
Weather Ankara	WAN	9	1609
Weather Izmir	WIZ	9	1461
Wisconsin Breast Cancer	WBC	32	194
Yacht Hydrodynamics	YH	6	308

regression datasets with different numbers of instances and variables (covering a range from 2 to 60 input variables and from 43 to 4177 examples). These datasets have been downloaded from the following web pages: UCI Machine Learning Repository[4], KEEL-dataset[5], Dataset Collections of Weka[6], and Luis Torgo Repository[7].

In all the experiments, we adopted a *5-fold cross-validation model*, i.e., we randomly split the dataset into 5-folds, each of which contains 20% of the examples of the dataset, and used four folds for training and one for testing. For each of the five partitions, we executed six trials of the algorithms (6 different seeds). The web page associated with this paper includes the datasets used in this study (the 5-fold cross-validation partitions), which can be found in a downloadable zip file (https://www.ugr.es/~ralcala/papers/ijfs21/Datasets.zip).

Methods considered for the experiments are briefly described and their main characteristics are summarized in Table 2 referring some related methods[8, 9,10]. TS $_{SP2-SI}$ and LING1 are two state-of-the-art methods used to obtain interpretable, simple and accurate linguistic FRBSs, i.e., they also optimize the semantic and complexity interpretability measures. Therefore, they are the most directly related or comparable methods from the recent literature as they were also designed to search for transparency.

FSmogfs^{e}+Tun^{e} is a MOEA used to embed the learning of the DB with the wrapper RB generation plus a post-processing stage to perform multiobjective evolutionary MFs with tuning and rule selection. It is a state-of-the-art pure linguistic accuracy-oriented method without further interpretability restrictions. To the best of our knowledge, this method obtains the most accurate pure linguistic FRBSs to date. Even though it should not be directly compared to interpretability oriented methods due to that fact that it is free to fully focus on accuracy, we have considered this method since it is actually representative of the expected accuracy of linguistic FRBSs. For a more detailed description of the methods, please refer to the references in Table 2.

The values of the parameters used to execute the proposed method at both stages are shown in Table 3. All of them are general parameters and not specific to any dataset, so that any user can set them (they are recommended standard parameters) without needing to adapt them to new problems. The values for the remaining algorithms are those proposed by the authors in the corresponding papers. We should note that all of them perform equitably in 100,000 evaluations.

In order to assess whether significant differences exist among the results, we adopt statistical analysis based on non-parametric tests, according to the recommendations made in [29] and [30], where a set of simple, safe, and robust non-parametric tests for statistical comparisons of classifiers has been introduced. In particular, we will use these non-parametric tests for a multiple comparison [29, 30]: Friedman's test and Holm's method. Moreover, we have applied Wilcoxon's signed ranks test for pair-wise comparison [29, 30]. A detailed description of these tests can be seen in http://sci2s.ugr.es/sicidm/.

5.2 Comparison of Methods Considering Semantic Interpretability Measures

In order to evaluate the effectiveness of the whole proposal, two state-of-the-art methods for obtaining interpretable, simple, and accurate linguistic FRBSs have been

[4] https://archive.ics.uci.edu/ml/index.php.

[5] https://sci2s.ugr.es/keel/datasets.php.

[6] https://cs.waikato.ac.nz/ml/Weka/datasets.html.

[7] http://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html.

[8] IntMeas: indicates whether the algorithm includes and optimizes interpretability measures.

[9] WM: Wang and Mendel algorithm [39].

[10] Tree-based RB generation: Proposed in Sect. 4.1.2.

Table 2 Methods considered in the comparisons

Method	Refs.	IntMeas	Description	Objectives
TS_{SP2-SI}	[19]	Yes	MOEA for Tuning and rule Selection (TS)	*MSE*/NR/GM3M
LING1	[21]	Yes	MOEA TS with L-IRL	*MSE*/NR/GM3M/RMI
FSmogfs^e+Tun^e	[35]	No	MOEA for embedded learning of DB with wrapper RB generation by WM + TS MOEA	*MSE*/NR
Proposed method	–	Yes	CFLTS-based MOEA for embedded learning of DB with wrapper linguistic Tree-based RB generation + TS MOEA	*MSE*/NR/GM3M/RMI

Table 3 Parameters

Parameter	Value
Mutation probability	$P_m = 0.2$
Size of the population	200
Size of the external population	61
Number of evaluations	100000 (in total for both stages)
Bits per gen	$BGene = 30$

considered for comparisons, i.e., they also optimize the semantic and complexity interpretability measures. They are: TS $_{SP2-SI}$ proposed in [19] and LING1 proposed in [21].

The results obtained by the studied methods are shown in Table 4. This table is grouped in columns by algorithms and it shows the average of the results obtained by each algorithm in all the studied datasets. For each algorithm, the first and second columns show both the average number of used variables and rules. The third column shows the average MSE in the test data (Tst.). The fourth and fifth columns show the interpretability measures GM3M and RMI. Finally, the last two rows of the table show the global average values (*Average*) and the percentage of worsening with respect to the proposed method (*% worsen*).

The results in this table show that the proposed method obtains the best *Average* results in all the analyzed interpretability measures. For all of them, the improvement with respect to the previous proposals is quite significant. Moreover, a significant number of datasets (in bold) can be observed for which the proposed approach obtains more accurate models.

Table 5 shows the rankings by carrying out Friedman's tests on the different methods considered in all the measures in this study (Var, NR, MSE_{tst}, GM3M and RMI). The best rankings are obtained by the proposed method in all the measures considered.

Moreover, Table 6 shows the adjusted p-values (apv) obtained using Holm's test, comparing all the methods versus the proposed method in all the measures. The results show that the proposed method outperforms all the methods in all the measures..

Finally, the web page associated with this paper (https://www.ugr.es/~ralcala/papers/ijfs21) shows some of the obtained linguistic models including two known test problems from the existent repositories, WAN (Weather in Ankara) and WBC (Wisconsin Breast Cancer) datasets, and analyses them from an interpretability point of view.

5.3 Comparison with a State-of-the-Art Pure Linguistic "Accuracy-Oriented" Method

In this section, a study of the proposed method compared to a method aimed at precision (FSmogfs^e+Tun^e) is carried out. FSmogfs^e+Tun^e [35] is a MOEA used for embedded learning of the DB with wrapper RB generation using WM[11], plus a post-processing stage to perform multiobjective evolutionary MF tuning and rule selection. Although it has these two objetives (*MSE* and *NR*), it actually uses them to try obtain more accurate solutions without further interpretability restrictions. To our knowledge, this method obtains the most accurate and pure linguistic FRBSs in regression problems to date. Although these problems are not comparable since our proposal is assumed to lose accuracy at some level in favor of transparency, we have included this comparison to show that it is still possible to improve both together, obtaining higher accuracy with much simple and more transparent models.

The results of the precision-oriented method are shown in Table 7. These results also include the values in GM3M and RMI measures (although the method does not originally consider these measures, we have computed them on the final models obtained by this method with comparative purposes). In Table 7, we use the same terminology as in the previous Table 4. The results in this table show how the proposed method once again obtains the best global average values (*Average*) for all the interpretability measures, where it also shows significant improvements. Moreover, it

[11] WM: Wang and Mendel algorithm [39].

Table 4 Results for comparison to methods including and optimizing semantic interpretability measures (best results for each dataset and metric in bold)

Datasets	TS_{SP2-SI}					LING1				
	Var	NR	MSE_{tst}	GM3M	RMI	Var	NR	MSE_{tst}	GM3M	RMI
ABA	8	16.27	2.513	0.450	0.692	8	13.00	2.780	0.470	0.660
ANA	7	88.17	0.006	0.204	0.154	7	155.27	0.006	**0.470**	0.000
BAS	16	83.37	389547	0.283	0.571	16	271.27	440395	0.411	0.584
BOS	13	173.33	13.957	0.325	0.384	13	13248.60	14.458	0.414	0.384
CPU	6	23.47	2246.11	0.395	0.564	6	31.60	2229.65	0.466	0.583
DIA	2	10.67	0.268	0.171	0.468	2	18.57	0.263	0.364	0.277
ELE	4	29.30	14851	0.528	0.504	4	32.50	18822	0.540	0.540
FAT	14	83.23	5.292	0.673	0.311	14	64.30	4.831	0.705	0.733
FOR	12	102.83	**2211**	0.215	0.734	12	507.47	4467	**0.473**	0.566
FRI	5	494.83	2.047	0.712	0.100	5	707.57	**1.933**	0.702	0.308
MOR	15	15.40	0.034	0.541	0.744	15	9.00	0.045	0.600	**0.970**
Mpg6	5	53.27	5.034	0.314	0.305	5	65.93	5.694	0.463	0.332
Mpg8	7	82.67	5.436	0.289	0.395	7	107.30	5.037	0.463	0.452
PRI	15	59.07	4104134	0.236	0.650	15	90.67	4816082	0.434	0.650
QUA	3	27.20	0.0182	0.275	0.512	3	96.73	0.0185	0.470	0.000
STP	9	32.87	**0.775**	0.365	0.513	9	14.10	1.640	0.470	0.760
STR	6	125.03	225851	0.644	0.332	6	174.17	248910	0.644	0.095
TRE	15	17.67	0.048	0.533	0.746	15	9.00	0.055	0.630	**0.980**
TRI	60	119.73	0.0133	0.119	0.568	60	170.37	0.0126	0.360	0.382
WAN	9	39.33	2.016	0.456	0.482	9	9.50	2.825	0.570	0.910
WIZ	9	29.20	1.095	0.493	0.487	9	13.00	1.525	0.610	0.930
WBC	32	142.73	948.41	0.218	0.754	32	246.567	928.64	0.397	0.710
YH	6	81.20	25.677	0.725	0.565	6	182.80	21.341	0.694	0.088
Average	12.09	83.95	–	0.398	0.502	12.09	140.84	–	0.514	0.517
% Worsen	510.10	1619.35	–	40.82	45.39	510.10	2784.48	–	23.67	43.69

Datasets	Proposed method				
	Var	NR	MSE_{tst}	GM3M	RMI
ABA	**1.2**	**3.67**	**2.456**	**0.777**	**0.980**
ANA	**1.0**	**3.93**	0.004	0.400	**0.894**
BAS	**2.0**	**5.73**	**243569**	**0.606**	**0.954**
BOS	**3.1**	**6.93**	**8.276**	**0.619**	**0.805**
CPU	**2.2**	**4.93**	**1739.53**	**0.700**	**0.935**
DIA	**1.7**	**3.50**	**0.179**	**0.611**	**0.916**
ELE	**2.0**	**5.37**	**12184**	**0.679**	**0.928**
FAT	**1.6**	**3.83**	**1.292**	**0.862**	**0.973**
FOR	**1.0**	**2.07**	2351	0.430	**0.826**
FRI	**2.9**	**6.80**	3.041	**0.778**	**0.956**
MOR	**3.1**	**6.00**	**0.017**	**0.696**	0.708
Mpg6	**2.0**	**5.40**	**4.524**	**0.717**	**0.947**
Mpg8	**2.3**	**6.07**	**4.224**	**0.647**	**0.960**
PRI	**2.4**	**5.90**	**2700226**	**0.683**	**0.961**
QUA	**1.0**	**1.63**	**0.0179**	**0.547**	**0.890**
STP	**2.7**	**5.67**	1.396	**0.708**	**0.962**
STR	**1.4**	**3.70**	**172589**	**0.702**	**0.906**
TRE	**1.9**	**4.67**	**0.038**	0.713	0.917
TRI	**2.7**	**5.73**	**0.0116**	**0.735**	**0.964**
WAN	**2.0**	**5.67**	**1.506**	**0.699**	**0.959**
WIZ	**2.3**	**5.30**	**1.015**	**0.729**	**0.981**
WBC	**1.4**	**4.27**	**747.16**	**0.615**	**0.956**
YH	**1.6**	**5.53**	**0.895**	**0.829**	**0.843**
Average	**1.98**	**4.88**	–	**0.673**	**0.918**
% Worsen					

Table 5 Rankings using Friedman's test on Var, NR, GM3M, RMI, and Tst

Algorithm	Rank. on Var	Rank. on NR	Rank. on GM3M	Rank. on RMI	Rank. on Tst
PROPOSED METHOD	1.0	1.00	1.09	1.13	1.17
TS $_{SP2-SI}$	2.5	2.30	2.89	2.52	2.24
LING1	2.5	2.70	2.02	2.35	2.59

Table 6 Adjusted p-values versus *Proposed Method* on Var, NR, GM3M, RMI, and Tst

Algorithm	apv on Var	apv on NR	apv on GM3M	apv on RMI	apv on Tst
TS $_{SP2-SI}$	7.28E–7	9.72E–6	1.88E–9	4.76E–6	3.03E–4
LING1	7.28E–7	1.78E–8	1.52E–3	3.65E–5	3.30E–6

Table 7 Results for comparison to a state-of-the-art pure linguistic but only accuracy-oriented method (best results for each dataset and metric in bold)

Datasets	FSmogfs^e+Tun^e					Proposed method				
	Var	NR	MSE_{tst}	GM3M	RMI	Var	NR	MSE_{tst}	GM3M	RMI
ABA	3.0	8.00	2.509	0.326	0.316	**1.2**	**3.67**	**2.456**	**0.777**	**0.980**
ANA	3.0	10.13	**0.003**	0.244	0.319	**1.0**	**3.93**	0.004	**0.400**	**0.894**
BAS	6.0	16.60	261323	0.212	0.450	**2.0**	**5.73**	**243569**	**0.606**	**0.954**
BOS	4.6	21.03	9.909	0.187	0.319	**3.1**	**6.93**	**8.276**	**0.619**	**0.805**
CPU	4.2	15.57	2390.30	0.281	0.576	**2.2**	**4.93**	**1739.53**	**0.700**	**0.935**
DIA	2	11.63	0.261	0.197	0.430	**1.7**	**3.5**	**0.179**	**0.611**	**0.916**
ELE	**2.0**	8.00	**10548**	0.436	0.559	**2.0**	**5.37**	12184	**0.679**	**0.928**
FAT	2.2	8.53	1.378	0.749	0.712	**1.6**	**3.83**	**1.292**	**0.862**	**0.973**
FOR	3.0	10.00	2628	0.166	0.728	**1.0**	**2.07**	**2351**	**0.430**	**0.826**
FRI	3.1	22.03	3.138	0.754	0.314	**2.9**	**6.80**	**3.041**	**0.778**	**0.956**
MOR	**2.0**	7.00	0.019	0.419	0.612	3.1	**6.00**	**0.017**	**0.696**	**0.708**
Mpg6	3.0	20.00	4.562	0.173	0.319	**2.0**	**5.40**	**4.524**	**0.717**	**0.947**
Mpg8	3.0	23.00	4.747	0.199	0.280	**2.3**	**6.07**	**4.224**	**0.647**	**0.960**
PRI	5.3	24.03	3344230	0.242	0.459	**2.4**	**5.9**	**2700226**	**0.683**	**0.961**
QUA	1.3	3.23	**0.0178**	0.091	0.668	**1.0**	**1.63**	0.0179	**0.547**	**0.890**
STP	3.0	23.00	**0.912**	0.197	0.307	**2.7**	**5.67**	1.396	**0.708**	**0.962**
STR	3.6	19.77	187917	0.667	0.422	**1.4**	**3.70**	**172589**	**0.702**	**0.906**
TRE	3.0	9.00	0.044	0.396	0.672	**1.9**	**4.67**	**0.038**	**0.713**	**0.917**
TRI	9.5	28.00	0.0119	0.179	0.393	**2.7**	**5.73**	**0.0116**	**0.735**	**0.964**
WAN	2.1	8.13	1.635	0.434	0.641	**2.0**	**5.67**	**1.506**	**0.699**	**0.959**
WIZ	**2.0**	10.00	**1.011**	0.370	0.607	2.3	**5.30**	1.015	**0.729**	**0.981**
WBC	5.5	25.50	766.89	0.180	0.185	**1.4**	**4.27**	**747.16**	**0.615**	**0.956**
YH	2.2	11.40	1.500	0.750	0.392	**1.6**	**5.53**	**0.895**	**0.829**	**0.843**
Average	3.42	14.94	–	0.341	0.464	**2.0**	**4.88**	–	**0.673**	**0.918**
% Worsen	72.71	205.97		49.29	49.44					

also gives more accurate results (in bold) in a good number of datasets..

Table 8 shows the results of the Wilcoxon's test for the proposed method and FSmogfs^e+Tun^e. The results show that the proposed method outperforms FSmogfs^e+Tun^e on Var, NR, GM3M, RMI, and Tst.

Additionally, on the web page associated with this paper (https://www.ugr.es/~ralcala/papers/ijfs21), we have also included an statistical comparison with some highly accurate general purpose state-of-the-art algorithms (such as Model Trees, Neural Networks, Random Forests, and Support Vector Machines that are available in recognized software tools) in order to create a simple benchmarking

Table 8 Wilcoxon's test to compare PROPOSED METHOD (Prop.) and FSMOGFS ^{e}+TUN ^{e} ([35]) (R^+) on Var, NR, GM3M, RMI, and MSE_{tst}

Methods analyzed	Comparison	R^+	R^-	Hypothesis ($\alpha = 0.05$)	p-value	% Improv.
Prop. vs. [35]	Var	262	14	Rejected	1.72E–4	72.71
Prop. vs. [35]	NR	276	0	Rejected	2.38E–7	205.97
Prop. vs. [35]	GM3M	276	0	Rejected	2.38E–7	49.29
Prop. vs. [35]	RMI	276	0	Rejected	2.89E–5	49.44
Prop. vs. [35]	Tst	220	56	Rejected	0.011	–

reference. Even though accuracy is not the main focus of the paper and that, of course, these approaches are definitely more accurate than our proposal, we just would like to show that we also achieve a really competitive performance from an accuracy point of view, while we usually get less than 5 linguistic rules.

5.4 A Case Study of a Real Problem Related to Childhood Obesity

This section includes a real example of the application of the proposed algorithm. This is a case study related to childhood obesity, where the analysis of the obtained model shown below has been provided by a real expert. Children being overweight and obese is a serious worldwide issue and one of the major health challenges in the twenty-first century [40]. Many children who are overweight/obese before puberty become obese in early adulthood.

Insulin resistance (IR), a reduced physiological response of the peripheral tissues to normal levels of insulin, is a growing concern that can result from childhood obesity [41]. Among all childhood obesity comorbidities, IR is the one that better correlates with the appearance of adverse cardiometabolic events in later life, including in particular, type 2 diabetes mellitus (T2DM) and cardiovascular disease (CVD) [42]. Several risk factors for IR in children have been suggested, e.g., body mass index (BMI) [43], central and peripheral adiposity, dietary factors, and physical activity (PA) [44].

Here, we present a dataset derived from a cohort of 1014 Spanish children ranging from 5 to 15 years, grouped in three experimental conditions (normal-weight, overweight and children with obesity). The study population is composed of 525 subjects in the obesity group, 201 in the overweight group, and 288 in the normal-weight group. For each group, a wide range of clinical and molecular data is available, including up to 850.000 genetic and epigenetic markers, more than 50 anthropometric and biochemical measurements, as well as lifestyle and physical activity (PA) data (obtained by means of food frequency questionnaires (FFQs) and accelerometers). As a particular case of study, we aimed to investigate the relationship between anthropometry, PA and IR status in the presented dataset of the children, which is representative of the typical complex biological records usually faced in life sciences. For this purpose, we selected a subsample of the aforementioned population consisting of 460 individuals presenting PA data of sufficient quality. The HOMA-IR index (Homeostasis Model Assessment for Insulin Resistance) was modeled as the outcome variable since it has been extensively validated as a good indicator of the IR status in children and adults [45, 46]. As predictor variables in the model, we employed clinical and anthropometric data (such as sex, age, puberty, height, waist circumference, and BMI), the main PA measurements (e.g., sedentary time (ST) and light, moderate, and vigorous PA), as well as biochemical indicators of cardiometabolic dysfunction that differ from HOMA-IR in obese children (e.g., HDL and LDL-cholesterol, triglycerides and two types of blood pressure). Unlike other datasets that suffer limitations due to the use of self-reported PA and ST measures, we used more objective PA measures like accelerometry, which allowed us to reduce study bias.

The results in Table 9 were obtained using a 5-fold cross-validation and show the good performance of the proposed method in this real problem. It obtains better results in all measures as compared with the remaining methods. Figure 9 includes an example of the linguistic models obtained. Unlike the examples shown on the web page (https://www.ugr.es/~ralcala/papers/ijfs21) associated with this work, which was analyzed without real knowledge of the problems, the analysis shown below of the model obtained has been provided by a real expert. The variables in this figure are ordered to represent the same order of splits in the tree generated when learning the rules. In this way, we can consider each split as a way to recognize the different divisions in the data from more general

Table 9 Results in the childhood obesity problem (best results for each metric in bold)

Methods	Var	NR	MSE_{tst}	GM3M	RMI
Proposed method	**1.83**	**4.37**	**0.911**	**0.716**	**0.956**
FSmogfs^{e}+Tun^{e}	3.73	12.30	0.950	0.242	0.425
TS_{SP2-SI}	15.00	320.37	1.430	0.314	0.649
LING1	15.00	590.97	1.450	0.377	0.641

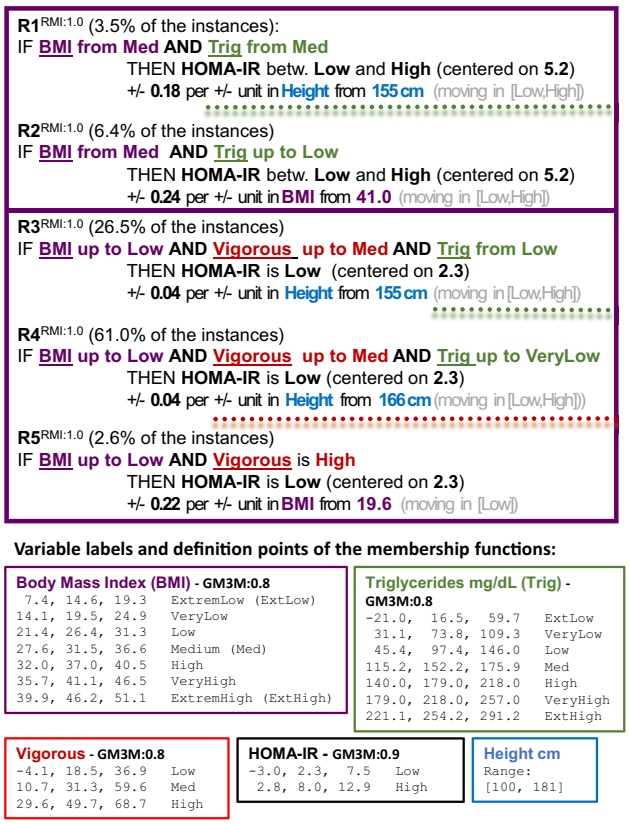

Fig. 9 KB obtained in the obesity problem (MSE_{tst}=0.706)

to more specific. We have used colors to ease the recognition of the different cases represented by the rules (same color per variable and split). Gray texts are included only to provide additional information, but this information is actually not a part of the rule structure proposed (and therefore, it is not needed for inference or for understanding). This is the same for the percentage of covered instances, for the GM3M and for the RMI values, since they are purely informative and refer to the semantic quality of each partition and rule, respectively.

As we previously explained, RMI goes from 1.0 (which represents that what a single rule asserts in its main coverage region is equal to what the model produces) to 0.0. We can see that all the rules are qualified with RMI equal to 1.0 and the variables with GM3M values over 0.8 (i.e., almost all of them can be considered to be strong equidistant linguistic partitions), so that the semantics are fully preserved. Therefore, in this case, the proposed method was able to properly describe the general behavior of the dataset and to identify the main relationships between predictor variables. In addition, the cut-off values proposed by the method for variable binning (or fuzzy discretization) were not only representative of the real variable domains but also from the clinical point of view.

The main relationship detected by the model was between the HOMA-IR and the BMI status. The method first suggested dividing the population into two groups of children according to their BMI values ($\sim >$ or $\sim <$ than 31). In this regard, rules generated in Fig. 9 show how children with a BMI of approximately less than 31 also have lower HOMA-IR values (and vice versa). Interestingly, and in agreement with this result, BMI has been previously presented as a risk factor for IR in child populations [43]. Within the group of children with a medium to low BMI, in turn, the method proposed additional divisions (rules) based on the level of vigorous PA. Children with higher levels of vigorous PA per day have the lowest HOMA-IR values. Interestingly, this finding is in line with previous studies, in which moderate-to-vigorous intensity PA has been inversely associated with cardiometabolic risk in children [47]. On the other hand, although it may be surprising that the method has not identified the ST variable as an important predictor of HOMA-IR, this fact is also in line with previous reports in which sedentary time has not been related as a risk factor when moderate-to-vigorous PA is taken into account [47]. In conclusion, the method proposes a set of easily understandable and clinically and biologically consistent rules, thus demonstrating a good performance on complex biological datasets. The rules generated also reveal behaviors and relationships between variables that are in line with previous findings reported with traditional statistics. Applying the method to genetics and environmental data together in future case studies could reveal new insights and point the way toward novel therapeutic targets for more precise interventions in childhood obesity.

6 Conclusions

This contribution is focused on the ability of the models obtained to explain regression problem with two main motivations:

- Understand and analyze a part of the underlying available data in order to check that we have coherent data that support information that is already known.
- Understand and analyze another part of the underlying data available to also explain unknown behaviors and relationships between variables, or to discover unknown, interesting and useful information.

In this study, in order to describe this information in a way that most resembles human expressions, we proposed an extension of the grammar (based on the composition of simple linguistic terms, including linguistic terms that are more general and resemble the way that humans speak, namely CFLTSs) and the fuzzy linguistic rule structure. It

allows both the general behavior of the data and its specific variability to be expressed in the same rule, so that in general fewer rules are needed to learn accurate linguistic models. However, the major contribution of this paper is the proposal of a novel interpretable linear extension of the consequent rule structure. This is of great importance for regression as it is a key point to maintain competitive (or even improved) accuracy. The consequent rule structure has been extended to explain the specific variability of the rule by means of two parameters that explain simple linear relations that can still be interpreted linguistically. Furthermore, we have proposed a method in two stages (learning DB+RB and tuning with rule selection) to optimize accuracy together with some interpretability measures. The main contribution in this part is the use of a new linguistic RB tree-based learning that adapts perfectly to the new type of rules.

We have statistically tested the proposal on 23 regression datasets with different complexities. The results obtained show the effectiveness of the proposed method by applying the Holm's, Friedman, and Wilcoxon tests to all the interpretability indexes, including accuracy, as it outperforms some of the previous state-of-the-art methods for obtaining interpretable pure linguistic FRBSs and also outperforms (even in accuracy) a state-of-art accuracy-oriented method. One of the most remarkable advantages is that it usually obtains very simple models, with only 4 or 5 rules, while still improving accuracy with respect to much more complex linguistic models. Furthermore, additional and useful information has been obtained that has never been seen before in previous linguistic fuzzy proposals, such as relatively simple linear relations with linguistic interpretations.

Finally, we have included a representative example of the linguistic model obtained in a case study on a real problem related to childhood obesity, where the analysis shown on the obtained model is provided by a real expert. The descriptions obtained seem to be clear and coherent, as we have been able to interpret them easily and as indicated to us by the expert.

Acknowledgements This paper has been supported by the Andalusian Government under Grant P18-RT-2248, the Health Institute Carlos III/Spanish Ministry of Science, Innovation and Universities under Grant PI20/00711, and the Spanish Ministry of Economy and Competitiveness under Grant PID2019-107793GB-I00 and Grant PID2020-119478GB-I00.

References

1. Castelvecchi, D.: Can we open the black box of AI? Nature **538**, 20–23 (2016)
2. Knight, W.: The U.S. military wants its autonomous machines to explain themselves. MIT Technol. Rev. **1**, 16 (2017)
3. Gadd, S.: Computer system could kill rather than cure, doctors warn, The Copenhagen Post (2017). http://cphpost.dk/?p=92249
4. Montavon, G., Samek, W., Müller, K.-R.: Methods for interpreting and understanding deep neural networks. Digital Signal Process. **73**, 1–15 (2018). https://doi.org/10.1016/j.dsp.2017.10.011
5. Greene, D., Lauren Hoffmann, A., Stark, L.: Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In: Proceedings of the 52nd Hawaii International Conference on System Sciences, Grand Wailea, Maui, Hawaii, 2019, pp. 2122–2131. https://doi.org/10.24251/HICSS.2019.258
6. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation''. AI Mag. **38**(3), 50–57 (2017). https://doi.org/10.1609/aimag.v38i3.2741
7. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision making and a "right to explanation". In: ICMLWorkshop on Human Interpretability in Machine Learning (WHI), New York, NY, 2016, pp. 1–9
8. Council, A.U.P.P.: Statement on algorithmic transparency and accountability (2017). https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf
9. European Union Agency for Fundamental Rights: Handbook on European Data Protection Law, FRA. Publications Office of the European Union, Luxembourg (2018)
10. Gunning, D.: Explainable articial intelligence (XAI). tech. report, defen-se advanced research projects agency, Tech. rep., Arlington, DARPA-BAA-16-53 (2016)
11. Fernandez, A., Herrera, F., Cordon, O., Jose del Jesus, M., Marcelloni, F.: Evolutionary fuzzy systems for explainable artificial intelligence: why, when, what for, and where to? IEEE Comput. Intell. Mag. **14**(1), 69–81 (2019). https://doi.org/10.1109/MCI.2018.2881645
12. Alonso, J.M., Casalino, G.: Explainable artificial intelligence for human-centric data analysis in virtual learning environments. In: Burgos, D., Cimitile, M., Ducange, P., Pecori, R., Picerno, P., Raviolo, P., Stracke, C.M. (eds.) Higher Education Learning Methodologies and Technologies Online, pp. 125–138. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-31284-8_10
13. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. Inform. Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
14. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**, 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

15. Watson, D.S., Krutzinna, J., Bruce, I.N., Griffiths, C.E., McInnes, I.B., Barnes, M.R., Floridi, L.: Clinical applications of machine learning algorithms: beyond the black box. BMJ **364**, 32 (2019). https://doi.org/10.1136/bmj.l886
16. El-Sappagh, S., Alonso, J.M., Ali, F., Ali, A., Jang, J., Kwak, K.: An ontology-based interpretable fuzzy decision support system for diabetes diagnosis. IEEE Access **6**, 37371–37394 (2018)
17. Song, X., Qin, B., Xiao, F.: FR-KDE: a hybrid fuzzy rule-based information fusion method with its application in biomedical classification. Int. J. Fuzzy Syst. **23**, 392–404 (2021). https://doi.org/10.1007/s40815-020-00957-z
18. Xu, C., Qian, G., Wang, H.: Stochastic multiple criteria comprehensive evaluation based on probabilistic linguistic preference relations: a case study of healthcare insurance audits in china. Int. J. Fuzzy Syst. **22**, 1607–1623 (2020). https://doi.org/10.1007/s40815-020-00865-2
19. Gacto, M.J., Alcalá, R., Herrera, F.: Integration of an index to preserve the semantic interpretability in the multi-objective evolutionary rule selection and tuning of linguistic fuzzy systems. IEEE Trans. Fuzzy Syst. **18**(3), 515–531 (2010). https://doi.org/10.1109/TFUZZ.2010.2041008
20. Gacto, M.J., Alcalá, R., Herrera, F.: Interpretability of linguistic fuzzy rule-based systems: an overview of interpretability measures. Inf. Sci. **181**(20), 4340–4360 (2011). https://doi.org/10.1016/j.ins.2011.02.021
21. Galende, M., Gacto, M.J., Sainz, G., Alcalá, R.: Comparison and design of interpretable linguistic vs. scatter FRBSs: GM3M generalization and new rule meaning index (RMI) for global assessment and local pseudo-linguistic representation. Inform. Sci. **282**, 190–213 (2014). https://doi.org/10.1016/j.ins.2014.05.023
22. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**(2), 81–97 (1956). https://doi.org/10.1037/h0043158
23. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning, parts I, II and III. Inform. Sci. **199**, 301–357 (1975)
24. Fernandez, A., Río, S., Herrera, F.: Revisiting evolutionary fuzzy systems: taxonomy, applications, new trends and challenges. Adv. Data Anal. Classif. **11**, 711–730 (2017)
25. Alcalá, R., Gacto, M., Alcalá-Fdez, J.: Evolutionary data mining and applications: a revision on the most cited papers from the last 10 years (2007–2017). Wiley Interdiscip. Rev. **8**(2), 1–17 (2018)
26. Fazzolari, M., Alcalá, R., Nojima, Y., Ishibuchi, H., Herrera, F.: A review of the application of multi-objective evolutionary systems: current status and further directions. IEEE Trans. Fuzzy Syst. **21**(1), 45–65 (2013)
27. Santiago, A., Dorronsoro, B., Nebro, A.J., Durillo, J.J., Castillo, O., Fraire, H.J.: A novel multi-objective evolutionary algorithm with fuzzy logic based adaptive selection of operators: fame. Inform. Sci. **471**, 233–251 (2019). https://doi.org/10.1016/j.ins.2018.09.005
28. Guillaume, S., Magdalena, L.: An OR and NOT implementation that improves linguistic rule interpretability. In: 11th World Congress of International Fuzzy Systems Association (IFSA 2005), Beijing (2005), pp. 88–92
29. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
30. García, S., Herrera, F.: An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. J. Mach. Learn. Res. **9**, 2677–2694 (2008)
31. Magdalena, L.: Semantic interpretability in hierarchical fuzzy systems: creating semantically decouplable hierarchies. Inform. Sci. **496**, 109–123 (2019). https://doi.org/10.1016/j.ins.2019.05.016
32. Cordon, O., Herrera, F.: A proposal for improving the accuracy of linguistic modeling. IEEE Trans. Fuzzy Syst. **8**(3), 335–344 (2000). https://doi.org/10.1109/91.855921
33. Quinlan, R.J.: Learning with continuous classes. In: Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, (1992), pp. 343–348
34. Holmes, G., Hall, M., Frank, E.: Generating rule sets from model trees. In: Proceedings of the 12th Australian Joint Conference on Artificial Intelligence: Advanced Topics in Artificial Intelligence, Vol. 1747 of Lecture Notes on Computer Science, Springer, (1999), pp. 1–12
35. Alcalá, R., Gacto, M.J., Herrera, F.: A fast and scalable multi-objective genetic fuzzy system for linguistic fuzzy modeling in high-dimensional regression problems. IEEE Trans. Fuzzy Syst. **19**(4), 666–681 (2011). https://doi.org/10.1109/TFUZZ.2011.2131657
36. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimization, In: Proceeding of the Evolutionary Methods for Design, Optimization and Control with Application to Industrial Problems, Barcelona, 2001, pp. 95–100
37. Eshelman, L.J.: The CHC adaptive search algorithm: how to have safe search when engaging in nontraditional genetic recombination. In: Rawlin, G. (ed.) Foundations of Genetic Algorithms, vol. 1, pp. 265–283. Morgan Kaufman, New York (1991)
38. Eshelman, L.J., Schaffer, J.D.: Real-coded genetic algorithms and interval-schemata. Found. Genetic Algorithms **2**, 187–202 (1993)
39. Wang, L., Mendel, J.: Generating fuzzy rules by learning from examples. IEEE Trans. Syst. Man Cybernet. **22**(6), 1414–1427 (1992)
40. T. G. . Obesity Collaborators, Health effects of overweight and obesity in 195 countries over 25 years, N Engl J Med 377 (1) (2017) 13–27. https://doi.org/10.1056/NEJMoa1614362
41. Lévy-Marchal, C., Arslanian, S.S., Cutfield, W.S., Sinaiko, A., Druet, C., Marcovecchio, M.L., Chiarelli, F.G.: Insulin resistance in children: consensus, perspective, and future directions. J. Clin. Endocrinol. Metab. **94**, 5189–5198 (2010). https://doi.org/10.1210/jc.2010-1047
42. Martin, B., Warram, J., Krolewski, A., Soeldner, J., Kahn, C., Martin, B., Bergman, R.: Role of glucose and insulin resistance in development of type 2 diabetes mellitus: results of a 25-year follow-up study. Lancet **2**(8825), 925–929 (1992). https://doi.org/10.1016/0140-6736(92)92814-V
43. Lee, J.M., Okumura, M.J., Davis, M.M., Herman, W.H., Gurney, J.G.: Prevalence and determinants of insulin resistance among U.S. adolescents: a population-based study. Diabetes Care **29**(11), 2427–2432 (2006). https://doi.org/10.2337/dc06-07
44. Leite, S.A., Monk, A.M., Upham, P.A., Chacra, A.R., Bergenstal, R.M.: Low cardiorespiratory fitness in people at risk for type 2 diabetes: early marker for insulin resistance. Diabetol. Metab. Syndrome **1**(1), 8 (2009). https://doi.org/10.1186/1758-5996-1-8
45. Matthews, D., Hosker, J., Rudenski, A., Naylor, B., Treacher, D., Turner, R.: Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia **28**(7), 412–419 (1985)
46. Gungor, N., Saad, R., Janosky, J., Arslanian, S.: Validation of surrogate estimates of insulin sensitivity and insulin secretion in children and adolescents. J. Pediatr. **144**(1), 47–55 (2004). https://doi.org/10.1016/j.jpeds.2003.09.045
47. Skrede, T., Stavnsbo, M., Aadland, E., Aadland, K.N., Anderssen, S.A., Resaland, G.K., Ekelund, U.: Moderate-to-vigorous physical activity, but not sedentary time, predicts changes in cardiometabolic risk factors in 10-y-old children: the active smarter kids study. Am. J. Clin. Nutr. **105**(6), 1391–1398 (2017). https://doi.org/10.3945/ajcn.116.150540

Carmen Biedma-Rdguez received the B.Sc. and M.Sc. degrees in computer science from the University of Granada, Granada, Spain, in 2018 and 2019, respectively, where he is currently working toward the Ph.D. degree in the Department of Computer Science and Artificial Intelligence. Her current research interests include multiobjective genetic algorithms and genetic fuzzy systems, particularly the learning/tuning of fuzzy systems for modeling and control with a good trade-off between accuracy and interpretability.

María José Gacto received the M.Sc. degree in computer science and the Ph.D. degree in computer science from the University of Granada, Spain, in 1999 and 2010, respectively. She is currently an Assistant Professor with the Department of Software Engineering, University of Granada. She is a member of the Intelligent Systems and Data Mining Research Group, Department of Computer Science, University of Jaén, Jaén, Spain. From 2008 to 2021, she was with the Department of Computer Science, University of Jaén, Jaén. She has authored more than 40 papers in international journals, book chapters, and conferences. She has worked on several research projects supported by the Spanish government and the European Union. Her research interests include multiobjective genetic algorithms and genetic fuzzy systems, particularly the learning/tuning of fuzzy systems for regression and control with a good trade-off between accuracy and interpretability, imbalanced regression, as well as fuzzy association rules.

Augusto Anguita-Ruiz holds a BSc in Biochemistry and Molecular Biology and a PhD in Nutritional biochemistry. He is specialized in the analysis of complex biological datasets such as those composed of clinical, omics, biochemical, and environmental data. For the identification of early-life predictive and prognostic biomarkers in Obesity and T2D, he develops new analysis pipelines and implements existing algorithms able to handle complex omics data (genetics, epigenetics, transcriptomic). His main technical skills include a strong statistical, programming, and data visualization background, with special emphasis in the use of machine learning models. Summary at: https://youtu.be/dCYZ6xY69Z4. At the moment of the publication of this article, he is part of the Barcelona Institute for Global Health (ISGlobal), where he works as a post-doctoral research fellow. His main research tasks focus in the development of a toolbox of advanced, next-generation, exposome tools and a prospective exposome cohort, which will be used to systematically quantify the effects of a wide range of community-level and individual-level environmental risk factors on mental, cardiometabolic, and respiratory health outcomes.

Jesús Alcalá-Fdez received the M.Sc. and Ph.D. degrees in computer science from the University of Granada, Spain in 2002 and 2006, respectively. He is an Associate Professor in the Department of Computer Science and Artificial Intelligence, University of Granada. He has published more than 90 papers in international journals, book chapters, and conferences. He has worked on several research projects supported by the Spanish government, the Andalusian government, and business. His current research interests include data mining, fuzzy association rules, genetic fuzzy systems, multiobjective evolutionary algorithms, bioinformatics, and data mining software. Dr. Alcalá-Fdez acts as an Editorial Member of several international journals. He currently serves as vice-chair of the IEEE/CIS FSTC Task Force on Fuzzy Systems Software. He belongs to the list of researchers "World's Top 2% Scientists" provided by Stanford University in 2020 and 2021.

Rafael Alcalá received the M.Sc. degree in Computer Science in 1998 and the Ph.D. degree in Computer Science in 2003, both from the University of Granada, Spain. He is a Full Professor in the Dept. of Computer Science and A.I at the University of Granada. He has published over 100 papers in international journals, book chapters, and conferences. He currently serves as member of the editorial board of the IEEE Transactions on Fuzzy Systems international journal. He was a president of the FSTC "Genetic Fuzzy Systems" Task Force at the IEEE Computational Intelligence Society (2009-2014, vice president 2014-2018). He was a Program Co-Chair at GEFS 2010, Area Co-Chair at FUZZ-IEEE 2011 and General Co-Chair at GEFS 2011 and 2013. His current research interests include multiobjective genetic algorithms and genetic fuzzy systems, particularly the learning/tuning of fuzzy systems for regression and control with a good trade-off between accuracy and interpretability, imbalanced regression, as well as fuzzy association rules.