https://dx.doi.org/10.1109/FUZZ45933.2021.9494598

Descriptive Stability of Fuzzy Rule-Based Systems

Corrado Mencar
Department of Computer Science
University of Bari Aldo Moro
70125 Bari, Italy
Email: corrado.mencar@uniba.it

Ciro Castiello
Department of Computer Science
University of Bari Aldo Moro
70125 Bari, Italy
Email: ciro.castiello@uniba.it

***Abstract*—Fuzzy Rule-Based Systems (FRBSs) are endowed with a knowledge base that can be used to provide model and outcome explanations. Usually, FRBSs are acquired from data by applying some learning methods: it is expected that, when modeling the same phenomenon, the FRBSs resulting from the application of a learning method should provide almost the same explanations. This requires a stability in the description of the knowledge bases that can be evaluated through the proposed measure of Descriptive Stability. The measure has been applied on three methods for generating FRBSs based on three benchmark datasets. The results show that, under same settings, different methods may produce FRBSs with varying stability, which impacts on their ability to provide trustful explanations.**

***Index Terms*—Fuzzy Rule-Based System, Explainability, Descriptive Stability.**

I. Introduction

A. Motivation and state-of-the-art

The pervasiveness of machine learning models in our daily lives is increasing at an ever-expanding rate, and humans already have to rely on the judgments, support, suggestions, and decisions of artificial tools. This is due to the level of maturity reached by artificial intelligence methods (in particular, machine learning methods), along with technological advancements that have enabled the creation of increasingly sophisticated tools which are now able to provide answers in real time in many fields of application.

Although some machine learning models, such as those based on (deep) neural learning, are able to perform very well in terms of accuracy of their outcomes, there is a strong need to go beyond a mere acquisition of the suggested results. Additional elements are demanded which may contribute to the acceptance of those results by the human user. This is especially true in some application fields (medicine, business intelligence, avionics, etc.) which cannot admit blind support from black boxes, because of the criticality of the choices to be made. Reliance on technology is eased when it can be explained: this brings to focus interest on such concepts as explanation [1], trust [2], ethics commitment and privacy protection [3] which are currently objects of investigation under the global umbrella of Explainable Artificial Intelligence (XAI) [4], [5], [6].

Ribeiro et al. highlighted two types of trust [2]: a trust in prediction (focussed on a single outcome proposed by a model), and an even more compelling trust in a model as a whole. Users need to be confident that a model will perform well on data according to some metrics of interest. In common practice of computer science this kind of assessment is usually performed in terms of accuracy: such an evaluation is often carried out by considering multiple models (possibly coming from a session of cross-validation), thus providing an indication about the goodness of the method that led to the generation of those models.

However, when trust comes into play, some different directions should be explored, other than just accuracy. In this sense, an interesting factor to take into account is represented by the consistency in terms of explanations. Ideally, explanations produced by one method should be somewhat comparable, especially when coming from the analysis of similar phenomena. Measuring such a consistency could be useful also to compare the explainability power of different methods. The idea is to spot (and reject) the occurrence known in the social sciences under the name of “Rashomon effect”, i.e. the situation where a single event is accounted and/or described in a discrepant/contradictory way by several subjects [7]. The theme of *Explanation Invariance* has been already introduced in the research context of XAI [8], although it is often put into practice by still resorting to some accuracy-based or outcome-centred evaluations [9].

B. Our approach and rationale

Following the outlined research direction, we went a step further in the investigation of explanation invariance, bearing in mind that a profitable machine learning method should be able to elicit some insights on a specific phenomenon under study. In this sense, the XAI practitioner must be interested in the explanations related to the process of phenomenon understanding provided by a model, rather than catching some of its inner mechanisms leading to the final inference.

In the attempt to identify the prominent features embedding the explanatory setup of a machine learning model, we focussed on the realm of fuzzy rule-based systems (FRBSs) and we considered the very structure of the rule-bases as a suitable illustration of the phenomenon at hand. FRBSs are commonly built up from data by means of several learning methods. Then the derived models are assessed in terms of accuracy of their results; also, they are regarded as interpretable tools for knowledge representation which oppose the power of natural language expressiveness to the opacity of black box systems. When we turn to consider the evaluation of a learning

978-1-6654-4407-1/21/$31.00 © 2021 IEEE

method in terms of explanation capabilities, we are interested in assessing something different from mere accuracy and we propose to look at the stability of its descriptions. In case of learning methods devoted to the automatic construction of fuzzy rule bases, such descriptions are encoded in the structure of the overall machinery responsible of the final inference process, which is expected to preserve stability under some circumstances.

Broadly speaking, if we repeatedly employ a method to extract fuzzy rule bases from the analysis of data, we would expect to come up with a number of models whose knowledge bases are not too dissimilar. As long as it is safe to assume that the repeated application of a method is aimed at modeling the same phenomenon (e.g., the training sets remain consistent in terms of data distribution), that seems to be a reasonable demand for a method which is called to capture and explain through its knowledge bases the investigated phenomenon. On the contrary, if the obtained models happen to show remarkable dissimilarities, that would be a signal of unsettled explanatory capabilities—either due to the method or to the violation of the aforementioned assumption—even if the models may prove to be effective in terms of accuracy. Should the method be used as a support in critical contexts of application (e.g., medical diagnosis), such an occurrence would be detrimental to user trusting.

In this way, the common practice of cross-validation may be performed to investigate a method in terms of its descriptive stability (rather than accuracy), in order to assess its explanatory power in place of just performance capabilities. To achieve this aim, a purposely defined metric is needed to evaluate the descriptive stability in a quantitative way. Once such metric is available, it may be also applied across the board to assess and compare the different values of descriptive stability characterizing a set of learning methods, so that a preference may be expressed among them whenever trustworthy explanation is a major issue.

In the next Section we introduce a method oriented to formally derive a degree of descriptive stability to be evaluated on a set of fuzzy rule bases. The degree is recursively calculated by considering the similarities traceable among the involved knowledge bases and all of their intrinsic components: all the formal passages leading to the definition of the stability degree are illustrated in a top-down fashion. In Section III we detail the experimental session performed on FRBSs derived by three different methods. The illustrated experiments concern both an *intra*-method evaluation (devoted to assess the stability of a particular learning method) and an *inter*-methods evaluation (oriented to highlight differences in terms of descriptive stability among different learning methods). Section IV closes the paper with some remarks and hints for future work.

The code and data are available under CC BY-NC-SA licence at https://github.com/cmencar/descriptive_stability/releases/tag/1.0.

II. Methodology

The Descriptive Similarity $\textsc{ds}$ is based on the recursive aggregation of the similarity of several components. In this work, we use arithmetic mean to aggregate similarities because of its compensative property. However, more drastic aggregation (e.g., through t-norms) or weighted aggregation (e.g., through OWA) are also possible. The definition of all the components of $\textsc{ds}$ is oriented to guarantee symmetry and values in $[0,1]$ so that two components have similarity equal to 1 if and only if they are identical; the use of arithmetic mean ensures that two components have similarity equal to 0 if and only if they do not share any element.

Definition 1. Let $\mathbf{S}$ be a set of n FRBSs that are enumerated as $S_1, S_2, \ldots, S_n$. The $\textsc{ds}$ degree is defined as

$$\textsc{ds}\left(\mathbf{S}\right) = \frac{2}{n^2-n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\textsc{kbs}\left(S_i,S_j\right) \tag{1}$$

The definition ensures that the descriptive stability metric does not depend on the ordering of FRBS, which is only instrumental to ensure that each pair (S_i, S_j) is considered just once.

The $\textsc{ds}$ degree depends on the Knowledge Base Similarity degree $\textsc{kbs}$, which compares the knowledge bases of two FRBSs. A FRBS is also endowed with an inference engine (which defines the semantics of the logical operators as well as the aggregation of rules); in this paper we assume that all the FRBSs in $\mathbf{S}$ share the same inference engine, therefore it is not involved in the definition of similarity.

In order to define the function $\textsc{kbs}$, we take into account the structure of a FRBS, which can be identified as a pair $(\textsc{db}, \textsc{rb})$ being $\textsc{db}$ the Data Base and $\textsc{rb}$ the Rule Base.

Definition 2. Let S', S'' two FRBSs with Data Bases $\textsc{db}', \textsc{db}''$ and Rule Bases $\textsc{rb}', \textsc{rb}''$ respectively. The degree $\textsc{kbs}$ is defined as

$$\textsc{kbs}\left(S',S''\right) = \frac{\textsc{dbs}\left(\textsc{db}',\textsc{db}''\right)+\textsc{rbs}\left(\textsc{rb}',\textsc{rb}''\right)}{2}$$

where $\textsc{dbs}$ is the Data Base Similarity function and $\textsc{rbs}$ is the Rule Base Similarity function as defined in the following sections.

A. Data Base Similarity

The database of a FRBS can be defined as a set of Linguistic Variables (LVs), i.e.

$$\textsc{db} = \{\textsc{lv}_1, \ldots, \textsc{lv}_m\}$$

where each LV is defined as the tuple

$$\textsc{lv} = (X, T, U, G, \mu)$$

defined as follows:

- X is the name of the LV, i.e. a symbol;
- T is the set of the linguistic terms, i.e. a set of symbols;
- U is the universe of discourse of the LV
- G is a generative grammar of the term set T;

- μ is an interpretation function, mapping each term $t \in T$ to a fuzzy set $\mu_t : U \mapsto [0, 1]$.

As a convention, we will use the "dot" notation to access the attributes of a LV. For example, by $\textsc{lv}.X$ we will denote the name X of the linguistic variable $\textsc{lv}$.

In order to define the similarity of the databases of two FRBSs, it must be noticed that they may not share the same LVs, or they may have LVs with the same name but different terms, or same name, same term set but different interpretations. The evaluation of the similarity between two Data Bases should take into account all these possibilities.

As a preliminary step, we define the set of names of a Data Base $\textsc{db}$ as follows:

$$\textsc{dbn} = \{\textsc{lv}.X : \textsc{lv} \in \textsc{db}\}$$

Here, we are assuming that all the LVs of a Data Base have distinct names.

Definition 3. Let $\textsc{db}', \textsc{db}''$ be two Data Bases. The Data Base Similarity degree $\textsc{dbs}$ is defined as:

$$\textsc{dbs}\left(\textsc{db}', \textsc{db}''\right) = \frac{\textsc{dbns}\left(\textsc{db}', \textsc{db}''\right) + \textsc{dbts}\left(\textsc{db}', \textsc{db}''\right) + \textsc{dbis}\left(\textsc{db}', \textsc{db}''\right)}{3}$$

where $\textsc{dbns}$ is the Data Base Name Similarity, $\textsc{dbts}$ is the Data Base Term Similarity and $\textsc{dbis}$ is the Data Base Interpretation Similarity.

$\textsc{dbns}$ evaluates how much two FRBSs share the same names in the Data Base; usually, such names correspond to features in data, therefore, $\textsc{dbns}$ evaluates the similarity of FRBSs in terms of feature sharing. In consequence of this definition, $\textsc{dbts}$ and $\textsc{dbis}$ are computed only for the fraction of LVs that are shared between the two Data Bases.

Definition 4. Let $\textsc{db}', \textsc{db}''$ be two Data Bases. The degree $\textsc{dbns}$ is defined as:

$$\textsc{dbns}\left(\textsc{db}', \textsc{db}''\right) = \frac{|\textsc{dbn}' \cap \textsc{dbn}''|}{|\textsc{dbn}' \cup \textsc{dbn}''|}$$

where $\textsc{dbn}', \textsc{dbn}''$ are the sets of names of $\textsc{db}', \textsc{db}''$, respectively.

Definition 5. Let $\textsc{db}', \textsc{db}''$ be two Data Bases. The degree $\textsc{dbts}$ evaluates the similarity of terms within the variables shared in the Data Bases. It is defined as:

$$\textsc{dbts}\left(\textsc{db}', \textsc{db}''\right) = \operatorname*{avg}_{\substack{\textsc{lv}' \in \textsc{db}' \\ \textsc{lv}'' \in \textsc{db}'' \\ \textsc{lv}'.X = \textsc{lv}''.X}} \textsc{ts}\left(\textsc{lv}'.T, \textsc{lv}''.T\right)$$

where avg is the arithmetic averaging function, and $\textsc{ts}$ is the Term Similarity function.

As a first approach, the function $\textsc{ts}$ can be defined as a quantification of the similarity of the term sets by using the Jaccard index. More refined approaches could use the generative grammars $\textsc{lv}'.G$ and $\textsc{lv}.G''$ to take into account possible relations between terms; also, if terms are ordered in a known way, this ordering could be used to define $\textsc{ts}$. In this work, we use the simplest approach.

Definition 6. Let T', T'' two term sets. The Term Similarity degree $\textsc{ts}$ is defined as:

$$\textsc{ts}\left(T', T''\right) = \frac{|T' \cap T''|}{|T' \cup T''|}$$

The Data Base Interpretation Similarity $\textsc{dbis}$ is an aggregate measure of similarities of all interpretations of the same terms occurring in LVs with the same names of the Data Bases of two FRBSs.

Definition 7. Let $\textsc{db}$', $\textsc{db}$" be two Data Bases. The Data Base Interpretation Similarity degree $\textsc{dbis}$ is defined as:

$$\textsc{dbis}\left(\textsc{db}', \textsc{db}''\right) = \operatorname*{avg}_{\substack{\textsc{lv}' \in \textsc{db}' \\ \textsc{lv}'' \in \textsc{db}'' \\ \textsc{lv}'.X = \textsc{lv}''.X \\ t \in \textsc{lv}'.T \cap \textsc{lv}''.T}} \textsc{is}\left(\textsc{lv}'.\mu\left(t\right), \textsc{lv}''.\mu\left(t\right)\right)$$

where $\textsc{is}$ is the Interpretation Similarity between two fuzzy sets.

An interpretation μ_t of a term t in a LV is a fuzzy set defined on $\textsc{lv}.U$. In order to define the similarity of interpretations, we should take into account the possibility that the interpretations are defined on different universes of discourses. In fact, although it is safe to assume that LVs with the same name refer to the same feature in data, yet it may be the case that the procedure used for designing the FRBSs is strictly data-dependent and the universe of discourse of each LV has been defined based on the available dataset, which could be different for each FRBS. In any case, we may safely assume that both universes of discourse are either discrete or continuous.

Definition 8. Given two fuzzy sets μ', μ'' defined on the universes of discourse U', U'' respectively, the Interpretation Similarity $\textsc{is}$ is defined as:

$$\textsc{is}\left(\mu', \mu''\right) = \frac{\int_{U' \cap U''} \min\left\{\mu'\left(x\right), \mu''\left(x\right)\right\} dx}{\int_{U' \cup U''} \max\left\{\bar{\mu}'\left(x\right), \bar{\mu}''\left(x\right)\right\} dx}$$

if U', U'' are continuous, and:

$$\textsc{is}\left(\mu', \mu''\right) = \frac{\sum_{x \in U' \cap U''} \min\left\{\mu'\left(x\right), \mu''\left(x\right)\right\}}{\sum_{x \in U' \cup U''} \max\left\{\bar{\mu}'\left(x\right), \bar{\mu}''\left(x\right)\right\}}$$

if U', U'' are discrete, where:

$$\bar{\mu}'\left(x\right) = \begin{cases} \mu\left(x\right), & x \in U' \\ 0, & x \notin U' \end{cases}$$

and $\bar{\mu}''$ is defined accordingly.

For the sake of efficiency, integral could be replaced with a sampled sum if necessary.

B. Rule Base Similarity

The Rule Base (RB) of a FRBS is defined as a set of rules. Comparing the RBs of two different FRBSs is not trivial because the RBs may have different cardinality, and rules may not share the same structure.

Assuming the availability of a function RS for evaluating the similarity between rules, the problem of computing the similarity of two RBs can be translated into an unbalanced assignment problem, which can be solved by different techniques, including the Hungarian algorithm specifically extended for unbalanced assignment [10].

Definition 9. Let $\mathrm{RB}', \mathrm{RB}''$ two Rule Bases consisting of r', r'' rules respectively. The Rule Base Similarity degree RBS is defined as:

$$\mathrm{RBS}\left(\mathrm{RB}', \mathrm{RB}''\right) = \frac{\mathrm{RBASS}\left(\mathrm{RB}', \mathrm{RB}''\right)}{\max\left\{r', r''\right\}}$$

where:

$$\mathrm{RBASS}\left(\mathrm{RB}', \mathrm{RB}''\right) = \max \sum_{\substack{R' \in \mathrm{RB}' \\ R'' \in \mathrm{RB}''}} \mathrm{RS}\left(R', R''\right) \cdot a\left(R', R''\right)$$

subject to:

$$\begin{cases} \sum_{R' \in \mathrm{RB}'} a\left(R', R''\right) = 1, & \text{if } r'' \leq r' \\ \sum_{R'' \in \mathrm{RB}''} a\left(R', R''\right) = 1, & \text{if } r' < r'' \\ a\left(R', R''\right) \in \{0, 1\} & \forall R', R'' \end{cases}$$

The function a stands for an assignment function which maps rules of one RB to rules of the other RB. As an example, if RB' has r' rules and RB'' has $r'' \leq r'$ rules, the unbalanced assignment problem matches each rule in RB'' with a rule in RB'; eventually, $r' - r''$ rules in RB' are left unassigned. The normalizing function $\max\{r', r''\}$ ensures that RBS is bounded in $[0, 1]$.

To define Rule Similarity, the structure of a rule R can be represented as a pair (A, C) corresponding to the antecedent and the consequent of the rule.

Definition 10. Given two rules R', R'', the degree RS is defined as:

$$\mathrm{RS}\left(R', R''\right) = \frac{\mathrm{AS}\left(A', A''\right) + \mathrm{CS}\left(C', C''\right)}{2}$$

where AS is the Antecedent Similarity and CS is the Consequent Similarity.

The antecedent of a rule can be simply defined as a conjunction of elementary soft constraints of the form "$X\ \mathtt{is}\ t$" where X is the name of a LV and t is a term in the corresponding term set. More complex rules can be defined, using different logical operators, nested structures, etc. In this work, we only consider conjunctive rules, therefore an antecedent A can be represented as a set of soft constraints SC. In turn, each soft constraint can be represented as a pair (X, t). By convention, we will denote by $\mathrm{SC}.X$ the name of the LV and by $\mathrm{SC}.t$ the linguistic term of SC. Also, it is convenient to define the set of all LV names occurring in an antecedent:

$$\mathrm{ASN} = \{\mathrm{SC}.X : \mathrm{SC} \in A\}$$

We also assume that no pair of soft constraints in the antecedent has the same name.

Definition 11. Let A', A'' two antecedents. The Antecedent Similarity degree AS is defined as:

$$\begin{aligned} &\mathrm{AS}\left(A', A''\right) = \\ &\operatorname{avg}\left\{\frac{\left|\mathrm{ASN}' \cap \mathrm{ASN}''\right|}{\left|\mathrm{ASN}' \cup \mathrm{ASN}''\right|}\right\} \cup \bigcup_{\substack{\mathrm{SC}' \in A' \\ \mathrm{SC}'' \in A'' \\ \mathrm{SC}'.X = \mathrm{SC}''.X}} \left\{\chi\left(\mathrm{SC}'.t, \mathrm{SC}''.t\right)\right\} \end{aligned}$$

where:

$$\chi\left(t', t''\right) = \begin{cases} 1 & t' = t'' \\ 0 & t' \neq t'' \end{cases}$$

The definition compares the terms of two soft constraints with the same name and returns 1 if the two terms coincide, otherwise it returns 0. More refined definitions may take into account additional properties coming from grammatical or ordering relations among terms.

The Consequent Similarity CS depends on the type of FRBS. In the case of a Mamdani FRBS, the consequent of a rule is a soft constraint, therefore the definition of similarity is similar to Antecedent Similarity.

Definition 12 (Mamdani)**.** Given two Mamdani consequents $C' = (Y', t')$ and $C'' = (Y'', t'')$, the Consequent Similarity degree CS is defined as:

$$\mathrm{CS}\left(C', C''\right) = \chi\left(C'.X, C''.X\right) \cdot \chi\left(C'.t, C''.t\right)$$

In the case of a 0-th order TSK FRBS, the consequent takes the form "$Y = w$" where Y is the name of a variable (not a LV) and $w \in \mathbb{R}$. In this case, the evaluation of similarity should take into account the metric properties of the output.

Definition 13 (0-th order TSK)**.** Given two 0-th order TSK consequents $C' = (Y', w')$ and $C'' = (Y'', w'')$, the Consequent Similarity (CS) degree is defined as:

$$\mathrm{CS}\left(C', C''\right) = \chi\left(C'.X, C''.X\right) \cdot e^{-\left(w' - w''\right)^2}$$

In the case of a classification FRBS (a.k.a. Fuzzy Rule Based Classifier, FRBC), the consequent takes the form of a class label, therefore the consequent similarity can be defined as in the Mamdani case.

III. Numerical Analysis and Discussion

To put into action the introduced methodology, we considered a number of methods designed to derive FRBCs from the analysis of raw data. Particularly, we selected three algorithms: i) the Fuzzy Unordered Rule Induction Algorithm (FURIA); ii) the Fuzzy Decision Tree algorithm (FDT); iii) a variant of FDT (FDT-S) oriented to simplify the produced knowledge bases.

TABLE I
PAIRWISE DS EVALUATIONS OF THE MODELS OBTAINED BY APPLYING FDT TO THE 10 FOLDS COMPOSING THE BEER DATASET.

	F-01	F-02	F-03	F-04	F-05	F-06	F-07	F-08	F-09	F-10
F-01	1									
F-02	0.771	1								
F-03	0.779	0.988	1							
F-04	0.777	0.985	0.996	1						
F-05	0.751	0.795	0.794	0.792	1					
F-06	0.781	0.986	0.997	0.994	0.797	1				
F-07	0.777	0.984	0.995	0.995	0.791	0.992	1			
F-08	0.780	0.989	0.999	0.995	0.794	0.996	0.994	1		
F-09	0.781	0.984	0.996	0.993	0.794	0.996	0.992	0.995	1	
F-10	0.736	0.926	0.929	0.932	0.752	0.928	0.929	0.930	0.928	1

The first method generates fuzzy rules especially focussing on accuracy, while disregarding the interpretability of the whole derived model [11]. FDT adopts a fuzzy version of the ID3 technique for tree generation to extract fuzzy rules from data viewed as fuzzy partitions [12]. The HILK procedure [13] can be employed to simplify a fuzzy knowledge base with the aim of increasing interpretability without penalizing accuracy too much: it has been adopted to realize a simplified variant of FDT (in the way it is implemented in GUAJE [14]) which stands as FDT-S, i.e. the third alternative method to learn fuzzy classifiers in our experimental session.

Three datasets have been selected to run the learning methods and derive the FRBCs to be evaluated: BEER [15], PIMA, and WINE [16]. A first evaluation has been performed on an *intra*-method basis: each of the three methods has been assessed in terms of its descriptive stability while tackling a single task. To do so, we performed a special case of 10-fold cross-validation, where the DS degree is involved in place of accuracy. The results of such an assessment are illustrated in Table I: for the sake of conciseness, the reported values of DS concern only the models obtained during the application of FDT on BEER. The matrix shows the degrees of the descriptive stability pairwise-evaluated for each couple of FRBCs extracted from the analysis of the 10 folds of data (F-01, . . ., F-10) composing the BEER dataset.

It can be observed how the maximum values of DS are obviously aligned on the diagonal of the matrix depicted in Table I, since they refer to the stability of a couple of identical models. The other values are in the range $[0.736, 0.999]$ and average to 0.902, thus testifying an adequate level of stability of FDT while describing the classification task underlying the BEER dataset. If we consider the columns of the matrix, we note how the 10 derived models are generally stable when pairwise considered, with the only exceptions of those pertaining to F-01 and F-05 whose degree of dissimilarity with respect to the others appears to be a little more pronounced.

We observe that the highest values of DS are obtained when the FRBSs differ in negligible details, such as small variations in the parameters defining the fuzzy sets involved in the linguistic variables. (FDT generates trapezoidal and triangular fuzzy sets.) As an example, FRBSs obtained in folds F-03 and F-08 (whose pairwise DS value amounts to 0.999) differ in the parameterization of 10 fuzzy sets, as reported in Table II. It should be noticed that DS does not explicitly depend on the parameters defining the fuzzy sets but it is more general since it uses the membership degrees to compute IS as in Definition 8. (In this case we approximate integrals with sampling, by dividing the domain of each LV in 100 equally spaced samples.)

TABLE II
DIFFERENT PARAMETERIZATION OF FUZZY SETS IN FRBSS GENERATED ON FOLDS F-08 AND F-03. TPZ=TRAPEZOIDAL FUZZY SET; TRN=TRIANGULAR FUZZY SET.

F-08	F-03
$\mathrm{tpz}(0.0, 0.0, 3.356, 9.917)$	$\mathrm{tpz}(0.0, 0.0, 3.367, 9.926)$
$\mathrm{trn}(3.356, 9.917, 17.12)$	$\mathrm{trn}(3.367, 9.926, 17.327)$
$\mathrm{trn}(9.917, 17.12, 27.561)$	$\mathrm{trn}(9.926, 17.327, 27.639)$
$\mathrm{trn}(17.12, 27.561, 35.387)$	$\mathrm{trn}(17.327, 27.639, 35.3)$
$\mathrm{trn}(27.561, 35.387, 40.636)$	$\mathrm{trn}(27.639, 35.3, 40.8)$
$\mathrm{tpz}(35.387, 40.636, 45.0, 45.0)$	$\mathrm{tpz}(35.3, 40.8, 45.0, 45.0)$
$\mathrm{tpz}(8.0, 8.0, 24.883, 46.291)$	$\mathrm{tpz}(8.0, 8.0, 25.304, 47.739)$
$\mathrm{trn}(24.883, 46.291, 91.841)$	$\mathrm{trn}(25.304, 47.739, 90.81)$
$\mathrm{trn}(46.291, 91.841, 145.667)$	$\mathrm{trn}(47.739, 90.81, 145.667)$
$\mathrm{trn}(91.841, 145.667, 201.6)$	$\mathrm{trn}(90.81, 145.667, 201.6)$

In the opposite case, we observe that the lowest value of DS (amounting to 0.736) is obtained when comparing the FRBSs of folds F-01 and F-10. In fact, the two FRBSs are different in many aspects, summarized as follows:

- different parameters in fuzzy sets, sometimes with significant differences (e.g., fuzzy sets $\mathrm{trn}(73.512, 101.976, 201.6)$ vs. $\mathrm{trn}(91.714, 145.667, 201.6)$ as interpretations of term HIGH for the LV BITTERNESS) for 17 fuzzy sets out of 25;
- different term sets (both in cardinality and terms) for the same LV (e.g., {VERY_LOW, LOW, AVERAGE_LOW, HIGH, VERY_HIGH} vs. {LOW, LOW_MEDIUM, MEDIUM_HIGH, HIGH} as term sets for LV BITTERNESS) for 3 LVs out of 4;
- different number of rules (29 vs. 24);
- different rules, in terms of number and values of soft constraints (e.g., "IF COLOR IS AMBER AND BITTERNESS IS LOW AND STRENGTH IS VERY HIGH THEN BEER_STYLE IS BARLEYWINE" VS. "IF COLOR IS AM-

TABLE III
GLOBAL DS EVALUATION OF THE THREE METHODS APPLIED TO THE THREE DATASETS.

	FDT	FDT-S	FURIA
BEER	0.902	0.850	0.760
PIMA	0.832	0.675	0.683
WINE	0.873	0.694	0.692

BER AND STRENGTH IS LABEL6 THEN BEER_STYLE IS BARLEYWINE") for 17 rules out of 24.

The global values of DS have been calculated to perform also an *inter*-methods evaluation involving all the three learning methods applied to the three datasets during the experimentation. The obtained values are included in Table III. We observe that FDT appears to be the most stable method. FURIA is by far the least stable method when applied on BEER, while being able to exhibit values of the DS degree similar to FDT-S on the remaining datasets. All in all, FURIA registers an average stability which is approximately 18% lower than FDT. To justify the superior stability of FDT over FDT-S we could infer that, being the latter characterized by the production of models which are simpler from a structural point of view, its evaluation in terms of DS is more influenced by (even reduced) alterations in the derived knowledge bases.

IV. CONCLUSIONS

The proposed Descriptive Stability metric is capable of evaluating the changes in the description of the knowledge bases of a set of FRBSs. It has been used as a stability measure to evaluate how much different are the knowledge bases of a set of FRBSs that have been generated with the same method under similar circumstances. We consider that this application assumes a particular relevance in XAI: it may be expected that, if the initial conditions for generating a model are almost the same, the resulting models should provide almost the same explanations. We propose the descriptive stability measure in the attempt to quantify this aspect and provide the designer with another tool for choosing a data driven method (other than accuracy). Indeed, preliminary experimental results show that methods can be significantly different in terms of descriptive stability.

The introduced measure is a first step for a more general tool to evaluate the descriptive stability of a number of models, including FRBSs with more complex rules, generative grammars for terms, and so on. The measure could be also extended with an ontology to grasp the semantic similarity of terms that are syntactically different (e.g., synonyms). Furthermore, it is possible to generalize the measure of descriptive stability by grasping the most general descriptive properties of predictive models, so as to assess and compare different methods generating predictive models with highly different structures (e.g., FRBSs, neural networks, generalized linear models, etc.). This is subject of ongoing research.

ACKNOWLEGMENT

The authors wish to express their thankfulness to Dr. José Alonso of the University of Santiago de Compostela, who provided the FRBSs employed in the experimental session. The research is partially supported by Ministero dello Sviluppo Economico (MISE) under grant F/190030/01-03/X44 "LIFT" and GNCS project "Computational Intelligence methods for Digital Health". The authors are members of the INdAM Research group GNCS and CILA (Centro Interdipartimentale di Logica e Applicazioni).

REFERENCES

[1] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2 2019.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?" - Explaining the Predictions of Any Classifier," in *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016, pp. 1135–1144.

[3] M. Taddeo and L. Floridi, "How AI can be a force for good," *Science*, vol. 361, no. 6404, pp. 751–752, 8 2018.

[4] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge-Based Systems*, vol. 8, no. 6, pp. 373–389, 12 1995.

[5] A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," *ArXiv*, vol. abs/2006.1, 2020.

[6] E. Lughofer, R. Richter, U. Neissl, W. Heidl, C. Eitzinger, and T. Radauer, "Explaining classifier decisions linguistically for stimulating and improving operators labeling behavior," *Inf. Sci.*, vol. 420, pp. 16–36, 12 2017.

[7] C. Davenport, "The Rashomon Effect, Observation, and Data Generation," in *Media Bias, Perspective, and State Repression*, 2012.

[8] K. Sokol and P. Flach, "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches," in *Proc. of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 56–67.

[9] K. K. Sethi, D. K. Mishra, and B. Mishra, "Novel algorithm to measure consistency between extracted models from big dataset and predicting applicability of rule extraction," in *2014 Conference on IT in Business, Industry and Government (CSIBIG)*. Institute of Electrical and Electronics Engineers Inc., 3 2014.

[10] L. Ramshaw and R. E. Tarjan, "On minimum-cost assignments in unbalanced bipartite graphs," *HP Labs, Palo Alto, CA, USA, Tech. Rep. HPL-2012-40R1*, 2012.

[11] J. Hühn and E. Hüllermeier, "FURIA: an algorithm for unordered fuzzy rule induction," *Data Mining and Knowledge Discovery*, vol. 19, no. 3, pp. 293–319, apr 2009.

[12] H. Ichihashi, T. Shirai, K. Nagasaka, and T. Miyoshi, "Neuro-fuzzy ID3: a method of inducing fuzzy decision trees with linear programming for maximizing entropy and an algebraic method for incremental learning," *Fuzzy Sets Syst.*, vol. 81, no. 1, pp. 157–167, 1996.

[13] J. M. Alonso and L. Magdalena, "HILK++: an interpretability-guided fuzzy modeling methodology for learning readable and comprehensible fuzzy rule-based classifiers," *Soft Computing*, pp. 1959–1980, 2011.

[14] J. M. Alonso, C. Castiello, L. Magdalena, and C. Mencar, *Explainable Fuzzy Systems - Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems*, ser. Studies in Computational Intelligence. Springer International Publishing, 2021, vol. 970. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-71098-9

[15] C. Castiello. (2017) BeerStyles3Features-1.0. [Online]. Available: http://dx.doi.org/10.17632/n4b6734rfn.1

[16] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255–287, 2011.