Integration of an Index to Preserve the Semantic Interpretability in the Multiobjective Evolutionary Rule Selection and Tuning of Linguistic Fuzzy Systems

María José Gacto, Rafael Alcalá, and Francisco Herrera

***Abstract*—In this paper, we propose an index that helps preserve the semantic interpretability of linguistic fuzzy models while a tuning of the membership functions (MFs) is performed. The proposed index is the aggregation of three metrics that preserve the original meanings of the MFs as much as possible while a tuning of their definition parameters is performed. Additionally, rule-selection mechanisms can be used to reduce the model complexity, which involves another important interpretability aspect. To this end, we propose a postprocessing multiobjective evolutionary algorithm that performs rule selection and tuning of fuzzy-rule-based systems with three objectives: accuracy, semantic interpretability maximization, and complexity minimization. We tested our approach on nine real-world regression datasets. In order to analyze the interaction between the fuzzy-rule-selection approach and the tuning approach, these are also individually proved in a multiobjective framework and compared with their respective single-objective counterparts. We compared the different approaches by applying nonparametric statistical tests for pairwise and multiple comparisons, taking into consideration three representative points from the obtained Pareto fronts in the case of the multiobjective-based approaches. Results confirm the effectiveness of our approach, and a wide range of solutions is obtained, which are not only more interpretable but are also more accurate.**

***Index Terms*—Fuzzy-rule-based systems (FRBSs), multiobjective evolutionary algorithms (MOEAs), rule selection, semantic interpretability index, tuning.**

I. Introduction

As DISCUSSED by Zadeh [1], computing with words (CW) is a methodology in which the objects of computation are words and propositions that are drawn from a natural language, e.g., small, large, far, etc. CW is inspired by the remarkable human capability to perform a wide variety of physical and mental tasks with no measurements or computations. CW-based techniques are employed to translate propositions that are expressed in a natural language into the generalized constraint language. The development of the methodology of CW is the development of a methodology in which words play the role of labels of perceptions. Linguistic variables and linguistic fuzzy rules are important elements in the conceptual structure of computational theory of perceptions (see [1, Fig. 4]). Further, as Zadeh stated [1], a fuzzy rule can be considered to be a Cartesian granule, and a fuzzy graph or a rule base (RB) may be viewed as a disjunction of Cartesian granules, and in essence, a fuzzy graph serves as an approximation to a function or a relation. This way, linguistic fuzzy modeling allows the modeling of systems to be dealt with by building a linguistic model that is interpretable by human beings. This task is usually developed by means of linguistic fuzzy-rule-based systems (FRBSs), which are also called Mamdani FRBSs [2], [3] and use fuzzy rules composed of linguistic variables [4]–[6] 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 represents the basis for the concept of linguistic *if–then* rules.

Many automatic techniques have been proposed to extract a proper set of linguistic fuzzy rules from numerical data. Most of them usually try to improve the performance that is associated with the prediction error without paying special attention to system's interpretability and without losing the linguistic meanings associated with the model. Finding the right interpretability–accuracy tradeoff, despite the original nature of fuzzy logic, has given rise to a growing interest in methods that take both aspects into account [7]–[11]. Ideally, both criteria should be satisfied to a high degree. However, since they are in conflict, this is not generally possible.

One way of doing this is to improve system's accuracy while trying to maintain interpretability to an acceptable level [9], [12]. By considering structural criteria, we can distinguish two main kinds of approaches that also take into account the interpretability of FRBSs.

1) *Complexity-based interpretability:* These approaches are used to decrease the complexity of the model that is obtained [12]–[21] (which are usually measured as the number of rules (NRs), variables, labels per rule, etc.).
2) *Semantics-based interpretability:* These approaches are used to preserve the semantics associated with the membership functions (MFs) [22]–[32]. We can find approaches that ensure semantic integrity, which usually

Manuscript received April 15, 2009; revised August 19, 2009 and November 13, 2009; accepted December 15, 2009. Date of publication January 19, 2010; date of current version May 25, 2010. This work was supported by the Spanish Ministry of Education and Science under Grant TIN2008-06681-C06-01.

M. J. Gacto is with the Department of Computer Science, University of Jaén, Jaén 23071, Spain (e-mail: mjgacto@ugr.es).

R. Alcalá and F. Herrera are with the Department of Computer Science and Artificial Intelligence, University of Granada, Granada 18071, Spain (e-mail: alcala@decsai.ugr.es; herrera@decsai.ugr.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2010.2041008

1063-6706/$26.00 © 2010 IEEE

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

imposes constraints on the MFs by considering measures such as distinguishability, coverage, fuzzy ordering, etc.

However, by paying attention to accuracy, one of the most widely used approaches to enhance the performance of FRBSs is the *tuning* of the MFs [27], [33]–[39]. It involves the improvement of a previous definition of the database (DB) once the RB has been obtained. The tuning methods refine the parameters that identify the MFs associated with the labels that comprise the DB [40]. Even though this approach is able to obtain highly accurate models, the semantic interpretability could be affected, depending on the variations that are performed in the MFs' shapes. The complexity of the models can also be a problem when a tuning is needed since usually, an excessive NRs is initially required to reach the highest degree of accuracy. Therefore, when an MF tuning is performed, three different criteria are required for a good accuracy–interpretability tradeoff: accuracy, complexity, and semantic interpretability.

A good way of optimizing these criteria simultaneously is the use of multiobjective evolutionary algorithms (MOEAs) [41], [42]. In fact, since this problem is multiobjective, most of the approaches that also take into account interpretability (especially, the complexity-based interpretability) use MOEAs to obtain a set of solutions with different degrees of accuracy and interpretability [13]–[15], [17]–[21], [23], [26].

In this paper, we propose an index to preserve the semantic interpretability of the DB while a tuning of the MFs is performed. The proposed index, i.e., GM3M, is defined as the geometric mean of three metrics, with the aim to minimize the displacement of the central point of the MFs, thus conserving the lateral amplitude rate of the MFs and maintaining the area of the original MFs that are associated with the linguistic labels. This measure can be used to quantify the interpretability of the tuned DB and could, therefore, be used as an objective within a multiobjective evolutionary process. To this end, we apply a specific MOEA to obtain interpretable and also accurate linguistic fuzzy models by concurrently performing a rule selection [16], [17], [43] and a tuning of the MF parameters with the following three objectives: *minimization of the system error*, *minimization of the NRs*, and *maximization of the proposed* Gm3m *index*. This postprocessing algorithm is based on the well-known *modified strength Pareto evolutionary algorithm* (*SPEA2*) [44]. It is called tuning and selection (TS) by SPEA2 for semantics-based index ($\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$). In order to improve its ability to search, $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$ implements such concepts as incest prevention and restarting [45] and incorporates the main ideas of the algorithm proposed in [13] to guide the search toward the desired Pareto zone. Thus, $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$ aims to generate a complete set of Pareto-optimum solutions, with different tradeoffs between accuracy and interpretability in the double sense, thus decreasing the complexity and maintaining the semantic-based interpretability. We have not considered the well-known nondominated sorting genetic algorithm version II (NSGA-II) [46] since, in [13], approaches based on SPEA2 were shown to be more effective when a tuning of the MFs is performed.

We tested our approach on nine real-world regression datasets. In order to analyze the interaction between the fuzzy rule selection and the tuning of MFs and how it can affect the different objectives, these are also individually proved in a multiobjective framework and compared with their respective single-objective counterparts [35]. We compared the different approaches by applying nonparametric statistical tests for pairwise and multiple comparisons [47]–[50] by considering three representative points from the obtained Pareto fronts in the case of the MOEAs. Results confirm the effectiveness of our approach, and a wide range of solutions is obtained, which are not only more interpretable but also more accurate.

Section II briefly analyzes the state of the art on interpretable linguistic FRBS modeling. Section III introduces the rule-selection and the tuning techniques, which are used concurrently in this paper. Section IV presents the proposed index to control the semantic interpretability of the MFs. Section V presents the $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$ algorithm and describes its main characteristics, as well as the considered genetic operators. Section VI shows the experimental study and the results obtained. Finally, in Section VII, we point out some conclusions. An Appendix has been included to describe the nonparametric tests that are used in our study.

II. Introduction to Interpretability on Linguistic Modeling

This section reviews some basic ideas and works on the linguistic modeling interpretability. Along with the review in [10], which widely represents most of the existing works in the specialized literature, a framework to categorize fuzzy model interpretability into high-level interpretability and low-level interpretability has been recently suggested in [11].

1) High-level interpretability is obtained on the fuzzy rule level by conducting overall complexity reduction in terms of some criteria, such as a moderate number of variables, a moderate NRs, completeness, and consistency of rules (complexity-based interpretability).
2) Low-level interpretability of fuzzy models is achieved on fuzzy set level by optimizing MFs in terms of the semantic criteria on MFs (semantics-based interpretability).

The complexity-reduction techniques that are used in traditional system modeling can serve as fuzzy rule optimization, which corresponds to aiming at the parsimony of the fuzzy RB, which is one of the main high-level interpretability criteria of fuzzy systems. This clarification is helpful as there are plentiful traditional system modeling methods on complexity reduction that have great potentials to induce compact RB in fuzzy system modeling. Earlier works [16], [17] used rule selection on an initial set of classification rules and two different criteria: accuracy and NRs. Along with the work presented in [17], Ishibuchi and coworkers [18]–[20] optimized such complexity criteria by applying MOEAs. Rule length (which is, sometimes, used in combination with the NRs) has been included to minimize the length of the rules by either rule selection [14], [18], [19] or rule learning [18], [20], [21]. A method has also been proposed in [13] and deeply discussed in [15] to minimize the NRs along with a tuning of the MFs.

Low-level interpretability is achieved by optimizing MFs on the fuzzy set level. Specifically, low-level interpretability hails

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

from the improvement on interpretability by introducing semantic constraint criteria into fuzzy modeling, which focus on the changes of MFs [11]. Classic approaches, such as [31] and [32], defined some helpful semantic criteria such as distinguishability, moderate number of MFs, natural zero positioning, normality, and coverage. These properties were later included in an MOEA to check their interaction when they evolve simultaneously [26]. Other works have focused on defining proper similarity metrics as a way to measure the distinguishability and coverage of the MFs [28], which are sometimes used to fix some minimum values of covering [24], [27], and some others are used to define maximum values of similarity for merging fuzzy sets and rules (particularly when MFs came from clustering techniques) [25], [30]. A similarity measure is also optimized in [29] to promote a good covering of the MFs, along with two complexity criteria in a combined index. Another MOEA is adopted in [23] to perform context adaptation. This algorithm considers the system error and an interpretability index to preserve the fuzzy ordering and a good distinguishability.

Additionally, some other works try to go a step ahead by considering all these kinds of measures in a linguistic framework in order to search for a more global definition of interpretability [12], [22]. In this sense, a conceptual framework is presented in [7] to characterize the interpretability of FRBSs. It makes reference to [10] and [11], which are combined in several interpretability levels (extending the low–high categorization).

Although most of the sematic-based approaches are mainly focused on finding partitions with a good overlapping among MFs (covering and distinguishability), in this paper, since interpretability is dependent on the problem context and user perceptions, we try to keep partitions and meanings to their original values, while performance improvements are still allowed. Further, it has also been combined with one of the classic complexity measures.

III. Fuzzy Rule Selection and Tuning of Membership Functions

In this paper, we present an MOEA for postprocessing that concurrently performs a fuzzy rule selection and a tuning of the MFs. This section briefly introduces the fuzzy-rule-selection technique and the tuning approach used to optimize the MF parameters.

A. Fuzzy Rule Selection

Fuzzy-rule-set-reduction techniques try to minimize the NRs of a given FRBS while maintaining (or even improving) the system's performance. To do this, erroneous and conflicting rules that degrade the performance are eliminated, thus obtaining a more cooperative fuzzy rule set and, as a result, potentially improving system's accuracy. Furthermore, in many cases, accuracy is not the only requirement of the model, but interpretability also becomes an important aspect. Reduction of the model complexity is a way to improve the system's readability, i.e., a compact system with few rules generally requires less effort in interpretation. Fuzzy-rule-set-reduction techniques are usually applied as a postprocessing stage once an initial fuzzy rule set has been extracted. One of the most used fuzzy-rule-set-reduction techniques is the rule selection. This approach involves obtaining an optimal subset of fuzzy rules from a previous fuzzy rule set by selecting some of them. We may find several methods for rule selection, with different search algorithms that look for the most successful combination of fuzzy rules [16], [17], [43]. An interesting heuristic rule-selection procedure is proposed in [51], where, by means of statistical measures, a relevance factor is computed for each fuzzy rule in the FRBSs to subsequently select the most relevant ones.

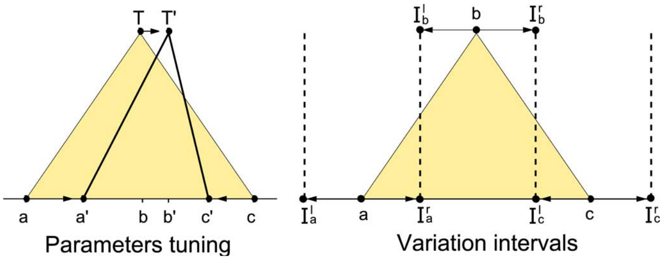

Fig. 1. Tuning by changing the basic MF parameters and the variation intervals.

These kinds of techniques for rule selection could be easily combined with other postprocessing techniques to obtain more compact and accurate FRBSs. This way, some works have considered the selection of rules along with the tuning of MFs by coding all of them (rules and parameters) in the same chromosome [13], [15], [33]–[35] within the same process and considering only performance criteria. Rules would be extracted only if it is possible to either maintain or even improve the system's accuracy. A very interesting conclusion from some of these recent works [15], [35] is that both techniques can present a positive synergy when they are combined within a well-designed optimization process.

B. Tuning of Membership Functions

This approach, which is usually called DB tuning, involves refining the MF shapes from a previous definition once the remaining FRBS components have been obtained [27], [36]–[39]. The classic way to refine the MFs is to change their definition parameters. For example, if the following triangular-shaped MF is considered:

$$\mu(x) = \begin{cases} \frac{x-a}{b-a}, & \text{if } a \leq x < b \\ \frac{c-x}{c-b}, & \text{if } b \leq x \leq c \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

changing the basic parameters—a, b, and c—will vary the shape of the fuzzy set that is associated with the MF, thus influencing the FRBS performance (see Fig. 1). This is also true for other shapes of MFs (trapezoidal, Gaussian, etc.).

Tuning involves fitting the characterization of the MFs associated with the primary linguistic terms that are considered in the system. Thus, the meaning of the linguistic terms is changed from a previous definition (i.e., an initial DB that is composed of the sematic concepts, and the corresponding MFs give meaning to them). As said, in order to preserve the semantic integrity

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

throughout the MF-optimization process [9], [31], [32], some researchers have proposed several properties. Considering one or more of these properties, several semantic constraints can be applied in the design process in order to obtain a DB that maintains the linguistic model integrity to the highest possible level [22], [24], [25], [29], [36].

In this paper, in order to illustrate the performance of the proposed approach, we use equidistributed strong fuzzy partitions [52] to define an initial set of triangular MFs. These kinds of fuzzy partitions, in which the sum of the membership degrees within the variable domain are equal to 1.0 and the triangular MFs are equidistant (therefore, symmetrical), perfectly meet the required semantic constraints, and they are widely assumed to have a high level of transparency. Anyhow, the initial DB should be given by an expert, if possible, since the concepts and their meaning strongly depend on the problem and the person who makes the assessment. In order to maintain the semantic integrity, we also consider some basic constraints by defining convenient variation intervals for each MF parameter. For each $\mathrm{MF}_j = (a_j, b_j, c_j)$, where $j = (1, \ldots, m)$, and *m* is the number of MFs in a given DB, the variation intervals are calculated as follows (see Fig. 1):

$$\begin{aligned}
[I_{a_j}^l, I_{a_j}^r] &= [a_j - (b_j - a_j)/2, a_j + (b_j - a_j)/2] \\
[I_{b_j}^l, I_{b_j}^r] &= [b_j - (b_j - a_j)/2, b_j + (c_j - b_j)/2] \\
[I_{c_j}^l, I_{c_j}^r] &= [c_j - (c_j - b_j)/2, c_j + (c_j - b_j)/2]. \qquad (2)
\end{aligned}$$

Due to these restrictions, it is possible to maintain the integrity of MFs to a reasonable level. In any case, it would be very interesting to have a measure for the quality of the tuned MFs. We propose three metrics that try to preserve the original form of the MFs, thus improving, if possible, the tradeoff between accuracy and interpretability.

IV. Semantic-Based Interpretability Index

In this section, we propose several metrics to measure the interpretability when a tuning is performed on the DB. At this point, we should remark that these metrics are based on the existence of the variation intervals (integrity constraints) that are defined in the previous section and, therefore, on the assumption that the initial DB comprises triangular MFs. Even though these measures and index are proposed to work with triangular MFs, they can be easily extended with some small changes in the formulation of Gaussian or trapezoidal MFs. Since significant changes in the DB can have a negative influence on interpretability, each metric is proposed to control how good some desirable aspects of the tuned MFs are with respect to the original ones (relative, not absolute, metrics). The metrics proposed are the following.

1) MFs displacement (δ): This metric measures the proximity of the central points of the MFs to the original ones. The closer they are to the original points, the higher the displacement.
2) MFs lateral amplitude rate (γ): This metric measures the left/right rate differences of the tuned and the original MFs. The closer the rates are, the higher the lateral amplitude rate.
3) MFs area similarity (ρ): This metric measures the area similarity of the tuned and the original MFs. It should be higher if the tuned and the original areas are closer.

In the following sections, the three proposed metrics will be explained in depth.

A. MFs Displacement Measure (δ)

This metric can control the displacements in the central point of the MFs. It is based on computation of the normalized distance between the central points of the tuned MF and the original MF, and is calculated by obtaining the maximum displacement attained on all the MFs. For each MF_j in the DB, we define $\delta_j = |b_j - b_j'|/I$, where $I = (I_{b_j}^r - I_{b_j}^l)/2$ represents the maximum variation for each central parameter. Thus, δ^* is defined as $\delta^* = \max_j\{\delta_j\}$. The δ^* metric takes values between 0 and 1; therefore, values near 1 show that the MFs present a great displacement. The following transformation is made so that this metric represents proximity (maximization):

$$\text{Maximize } \delta = 1 - \delta^*. \qquad (3)$$

This metric could also be used for either Gaussian or trapezoidal MFs by considering the middle of the core as the position to preserve.

B. MFs Lateral Amplitude Rate Measure (γ)

This metric can be used to control the shapes of the MFs. It is based on relating the left and right parts of the support of the original and the tuned MFs. Let us define left $S_j = |a_j - b_j|$ as the amplitude of the left part of the original MF support and right $S_j = |b_j - c_j|$ as the right-part amplitude. Let us define left $S_j' = |a_j' - b_j'|$ and right $S_j' = |b_j' - c_j'|$ as the corresponding parts in the tuned MFs. The variable γ_j is calculated using the following equation for each MF:

$$\gamma_j = \frac{\min\{\text{left } S_j/\text{right } S_j, \text{left } S_j'/\text{right } S_j'\}}{\max\{\text{left } S_j/\text{right } S_j, \text{left } S_j'/\text{right } S_j'\}}. \qquad (4)$$

Values near 1 mean that the left and right rates in the original MFs are highly maintained in the tuned MFs. Finally, γ is calculated by obtaining the minimum value of γ_j as

$$\text{Maximize } \gamma = \min_j\{\gamma_j\}. \qquad (5)$$

This metric always presents a value of 1 in the case of Gaussian MFs. It could also be used for trapezoidal MFs by considering the middle of the core as the central point, computing γ_j with the core extremes, computing γ_j with the MF extremes, and averaging both values.

C. MFs Area Similarity Measure (ρ)

This metric can be used to control the area of the shapes of the MFs. It is based on relating the areas of the original and the tuned MFs. Let us define A_j as the area of the triangle that represents the original MF_j and A_j' as the new area. The variable

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

ρ_j is calculated using the following equation for each MF:

$$\rho_j = \frac{\min\{A_j, A'_j\}}{\max\{A_j, A'_j\}}. \tag{6}$$

Values near 1 mean that the original area and the tuned area of the MFs are more similar (fewer changes). The ρ metric is calculated by obtaining the minimum value of ρ_j

$$\text{Maximize } \rho = \min_j\{\rho_j\}. \tag{7}$$

This metric is also applicable for trapezoidal and Gaussian MFs.

D. *Semantics-Based Interpretability Index Based on Aggregation of the Three Measures:* GM3M

We propose an aggregation of the metrics in a global index based on the geometric mean. As mentioned, this index is called GM3M and is defined as

$$\text{Maximize GM3M} = \sqrt[3]{\delta\gamma\rho}. \tag{8}$$

The value of GM3M ranges between 0 (which is the lowest level of interpretability) and 1 (which is the highest level of interpretability). The use of either $\min_j\{\cdot\}$ or $\max_j\{\cdot\}$ to compute the different metrics ensures the interpretability to a minimum level in all the MFs, since our main aim is to measure the worst case. Therefore, if there is a major problem in any of the MFs, it can be detected and reflected in each particular metric. Similarly, it is clear that if only one of the metrics has very low values, a problem arises in the interpretability. The used aggregation operator considers this fact. Moreover, all these relative metrics present complementary properties to measure the relation with the initial MFs.

V. Multiobjective Evolutionary Algorithm for Rule Selection and Tuning of Fuzzy Rule-Based Systems

Since it is not possible to either obtain the different interpretability–accuracy tradeoff degrees or handle the synergy of both approaches separately, the proposed algorithm performs a fuzzy rule selection along with a tuning of the MFs in order to improve the system's accuracy as a first objective, the model complexity as a second objective, and the GM3M index in order to preserve the semantic interpretability as the third objective. As mentioned, it is a specific MOEA that is called SPEA2 for semantic interpretability, i.e., $\text{TS}_{\text{SP2-SI}}$, which is based on the well-known SPEA2 [44] algorithm. In the next section, the main components of this algorithm are described, and then, the specific characteristics and its main steps are presented.

A. *Objectives*

Every chromosome is associated with a 3-D objective vector, each element of which expresses the fulfillment degree of the following three objectives:

1) semantic interpretability maximization: semantic-based index, GM3M;
2) complexity minimization: number of selected rules, NR;
3) error minimization: mean-squared error divided by 2 ($\text{MSE}_{/2}$).

The number of input variables is another complexity measure that could be considered to improve the system's interpretability. However, we have not used this measure since this can be considered in a previous stage, thus avoiding the use of a fourth objective in the MOEAs, which, nowadays, are not able to work properly with such quantity of objectives. The value of $\text{MSE}_{/2}$ of an FRBS that is decoded from a given chromosome is defined as follows: $\text{MSE}_{/2} = (1/2)|D|\sum_{l=1}^{|D|}(F(x^l) - y^l)^2$, where $|D|$ is the dataset size, $F(x^l)$ is the output of the FRBS when the lth example is an input, and y^l is the known desired output. The fuzzy inference system uses the *center of gravity weighted by the matching* strategy as a defuzzification operator and the *minimum t-norm* as implication and conjunctive operators.

B. *Coding Scheme and Initial Gene Pool*

A double coding scheme for both *rule selection* (C_{S}) and *tuning* (C_{T}) is used: $C^p = C^p_{\text{S}}C^p_{\text{T}}$. In the $C^p_{\text{S}} = (c_{\text{S}1}, \ldots, c_{\text{S}m})$ part, the coding scheme consists of binary-coded strings with size m (where m is the number of initial rules). Depending on whether a rule is selected or not, values of either "1" or "0" are, respectively, assigned to the corresponding gene. In the C_{T} part, a real coding is used, with m^i being the number of labels of each of the n variables in the DB

$$C^p_{\text{T}} = C_1 C_2 \ldots C_n$$
$$C_i = (a^i_1, b^i_1, c^i_1, \ldots, a^i_{m^i}, b^i_{m^i}, c^i_{m^i}), \qquad i = 1, \ldots, n.$$

The initial population is obtained with all individuals having all genes with value "1" in C_{S}. In the C_{T} part, the initial DB is included as a first individual, and the remaining individuals are generated at random within the corresponding variation intervals that are defined in Section III-B.

C. *Crossover and Mutation*

In this section, we propose an intelligent crossover and a mutation operator based on our experience in this problem. This is able to adequately profit from the parents when both rule selection and tuning are applied. The steps to obtain each offspring are as follows.

1) Blend crossover (BLX)-0.5 [53] is applied to obtain the C_{T} part of the offspring.
2) Once the offspring C_{T} part has been obtained, the binary part C_{S} is attained based on the C_{T} parts (MFs) of parents and offspring. For each gene in the C_{S} part that represents a concrete rule, the following hold.
   a) The MFs involved in such rule are extracted from the corresponding C_{T} parts for each individual that is involved in the crossover (offspring and parents 1 and 2). Thus, we can obtain the specific rules that each of the three individuals represent.
   b) Euclidean normalized distances are computed between the offspring rule and each parent rule by considering the center points (vertex) of the MFs that are composed of such rules. The differences

between each pair of centers are normalized by the amplitudes of their respective variation interval.

c) The parent with the rule closer to the one that is obtained by the offspring is the one that determines whether this rule is selected or not for the offspring by directly copying its value in C_{S} for the corresponding gene.

This process is repeated until all the C_{S} values are assigned for the offspring. Four offspring are obtained by repeating this process four times. (After considering mutation, only the two most accurate values are taken as descendants.) By applying this operator, exploration is performed in the C_{T} part, and C_{S} is directly obtained based on the previous knowledge that each parent has about the fact whether a specific configuration of MFs can be used for each rule. This avoids the possibility of recovering a bad rule that was discarded for a concrete configuration of MFs, while allowing the recovery of a good rule that is still considered for this concrete configuration, thus increasing the probability of success in either the selection or the elimination of a rule for each concrete configuration of MFs. Since a better exploration is performed for the C_{S} part, the mutation operator does not need to add rules. This way, once an offspring is generated, the mutation operator changes a gene value at random in the C_{T} part and directly sets to zero a gene that is selected at random in the C_{S} part (one gene is modified in each part) with probability P_m.

By applying these operators, two problems are solved. First, crossing individuals with very different rule configurations is more productive. Second, this way of working favors rule extraction since mutation is employed only to remove unnecessary rules.

D. Main Characteristics of TS$_{SP2\text{-}SI}$

The proposed algorithm uses the SPEA2-selection mechanism. However, in order to improve the algorithm's ability to search, the following changes are considered.

1) The proposed algorithm includes a mechanism for incest prevention based on the concepts of CHC [45] in order to avoid premature convergence in the C_{T} part (real coding), which is the main responsibility of accuracy improvements and represents a more complicated search space than the C_{S} part (binary coding). In CHC, only those parents are crossed whose Hamming distance divided by 4 is greater than a threshold. Since we consider a real coding scheme (i.e., only C_{T} parts are considered), we have to transform each gene using a gray code with a fixed number of bits per gene (BGene), which are determined by the system's expert. This way, the threshold value is initialized as $L = (\#C_{\mathrm{T}} \times \text{BGene})/4$, where $\#C_{\mathrm{T}}$ is the number of genes in the C_{T} part of the chromosome. At each generation of the algorithm, the threshold value decreases by 1, which allows crossing closer solutions. This mechanism can also be maintained because the parent selection is multiobjective, which provides a parent diversity that is similar to the original CHC.

2) The restarting operator forces the external population to be empty and generates a new initial population. This initial population includes a copy of the individuals with the best value in each objective (before removing them from the external population). The remaining individuals in the new population take the values of the most accurate individual in the C_{S} part and values generated at random in the C_{T} part. This preserves the most accurate and the most interpretable solutions that are obtained. The restarting operator is applied when we detect that all the crossovers are allowed. However, in order to avoid premature convergence, we apply the first restart if 50% of crossovers are detected at any generation (the required ratio can be defined as $\%_{\text{required}} = 0.5$). This condition is updated each time restarting is performed as $\%_{\text{required}} = (1+\%_{\text{required}})/2$. Moreover, the most accurate solution should be improved before each restart. To preserve a well-formed Pareto front, the restart is not applied at the end. The number of evaluations without restart can be estimated as the number of evaluations needed to apply the first restart multiplied by 10. Additionally, restart is disabled if it was never applied before reaching the midpoint of the total number of evaluations.

3) At each stage of the algorithm (between restarting points), the number of solutions in the external population ($\overline{P}_{t+1}$) that is considered to form the mating pool is progressively reduced, by focusing only on those with the best accuracy. To do this, the solutions are sorted from the best to the worst (considering accuracy as criterion), and the number of solutions that are considered for selection is reduced progressively from 100% at the beginning to 50% at the end of each stage. It is done by taking into account the value of L. In the last evaluations when restart is disabled, the mechanism to focus on the most accurate solutions (which is the most difficult objective) is also disabled to obtain a wide, well-formed Pareto front, from the most accurate solutions to the most interpretable ones.

The main steps of TS$_{\mathrm{SP2\text{-}SI}}$ are finally presented in Fig. 2 (see SPEA2 in [44]).

VI. Experimental Study

To evaluate the usefulness of the proposed approach, we used nine real-world problems. Table I summarizes the main characteristics of the nine datasets and shows the link to the knowledge extraction based on evolutionary learning (KEEL) software tool Web page (http://www.keel.es/) [55] from which these can be downloaded. This section is organized as follows.

1) Section VI-A presents the experimental setup.
2) Section VI-B analyzes the tuning of MFs individually, by paying attention to the GM3M index. To this end, the tuning component of the proposed approach and its single-objective counterpart are compared in terms of the most accurate solutions. Some example DBs are also presented in order to graphically show the effects of the use of the GM3M index as an objective in the evolutionary model.
3) Section VI-C presents an analysis on the rule selection individually. In order to better analyze the interaction between the different components of the proposed approach,

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

```
Input: $N$ (population size), $\overline{N}$ (external population size), $E$ (maximum number of evaluations),
    $BGene$ (bit per gene for gray code).
Output: $A$ (non-dominated set).
Terminology:
    $\#C_T$ (number of genes in the real part $C_T$), $\#O$ (number of objectives), $Evs$ (current number of evaluations),
    $L$ (threshold for incest prevention), $InitL = (\#C_T * BGene)/4$ (initial threshold), $R\%$ (descendant % required for restart), $Rst$ (variable to activate restart), $Nded$ (evaluations needed to form a Pareto),
    $Acc^+$ (accuracy improvement is detected in the most accurate solution from the latest restart).
Algorithm:
  1) Generate $P_0$ (initial population) and create $\overline{P}_0 = \emptyset$ (empty external population).
  2) Evaluate individuals in $P_0$ ($MSE_{/2}$) and set:
     – $L = InitL$; $R\% = 0.5$; $Rst = false$; $Evs = N$; $Nded = 0$; $t = 0$;
  3) Calculate fitness values of individuals in $P_t$ and $\overline{P}_t$. Copy all non-dominated individuals in $P_t \cup \overline{P}_t$ to $\overline{P}_{t+1}$. If $|\overline{P}_{t+1}| > \overline{N}$ apply truncation operator. If $|\overline{P}_{t+1}| < \overline{N}$ fill with dominated in $P_t \cup \overline{P}_t$.
  4) If $Evs \geq E$, return A and stop.
  5) If $(Rst)$ and $(Evs < E - Nded)$ and $(Acc^+)$:
     – $L = InitL$; $R\% = (R\% + 1)/2.0$; If $Nded$ is $0$, $Nded = Evs * 10$; $Rst = false$.
     – Copy the best individuals in each objective to $P_t$. Empty $\overline{P}_t$ ($\overline{P}_t = \emptyset$). Fill remaining $N - \#O$ individuals in $P_t$ with $C_T$ at random and $C_S$ equal to the most accurate individual.
     – Evaluate the $N - \#O$ new individuals in $P_t$ ($MSE_{/2}$), set $Evs += N - \#O$ and go to Step 3.
  6) Generate the next population:
     – If $Evs < E - Nded$, set $P = (L/(InitL * 2.0) + 0.5)$ else set $P = 1.0$. Perform binary tournament selection with replacement on the $\lfloor \overline{N} * P \rfloor$ most accurate solutions of $\overline{P}_{t+1}$ to fill the mating pool.
     – Apply crossover (BLX-Specific) and mutation for each two parents in the mating pool if the hamming distance between their $C_T$ part Gray codings divided by 4 is over $L$.
     – Set $P_{t+1}$ to the resulting population with the obtained $G$ descendant. Set $Evs += G * 2$.
  7) Variables updating:
     – If $L > 0$, $L = L - 1$; If $G \geq N * R\%$, $Rst = true$; If $Nded$ is $0$ and $Evs \geq E/2$, $Nded = E$.
  8) Go to Step 3 with $t = t + 1$.
```

Fig. 2. $\text{TS}_{\text{SP2-SI}}$ algorithm scheme.

the rule-selection component has also been compared with its single-objective counterpart.

4) Section VI-D analyzes the proposed approach and the interaction between the tuning and the rule-selection components. This analysis has been carried out in the same way, i.e., by considering only tuning and paying attention to the effects that the concurrent use of both techniques promotes to the different criteria, particularly to the GM3M index.
5) Section VI-E includes a global statistical analysis of the most accurate solutions by considering all the approaches and the corresponding optimized measures/objectives.
6) Finally, Section VI-F shows a graphical and statistical analysis of the obtained Pareto fronts. To perform this study, we plot the centroids (average values) of three representative points of the Pareto fronts (from the most accurate to the most interpretable) on the accuracy–complexity and accuracy–semantic planes. These plots provide a glimpse of the trend of the Pareto fronts. We also present a statistical analysis of the centroids of the most interpretable and intermediate solutions. For completeness, we also show some representative Pareto fronts that are achieved by the different MOEAs.

TABLE I
DATASETS THAT ARE CONSIDERED FOR THE EXPERIMENTAL STUDY

Datasets	**Name**	**Variables**	**Patterns**
Plastic Strength	PLA	3	1650
Quake	QUA	4	2178
Electrical Maintenance	ELE	5	1056
Abalone	ABA	9	4177
Stock prices	STP	10	950
Weather Ankara	WAN	10	1609
Weather Izmir	WIZ	10	1461
Mortgage	MOR	16	1049
Treasury	TRE	16	1049

Available at: http://sci2s.ugr.es/keel/datasets.php

A. Experimental Setup

In all the cases, the well-known *ad hoc* data-driven learning algorithm of Wang and Mendel [54] is applied to obtain an initial set of candidate linguistic rules. The initial linguistic partitions comprise *five linguistic terms* in the case of datasets with less than nine variables and *three linguistic terms* in the remaining ones (which helps obtain a more reasonable NRs in the more complex datasets). Once the initial RB is generated, the different postprocessing algorithms can be applied. The methods that are considered for the experiments are briefly described in Table II. In order to evaluate the advantages of concurrently performing rule selection and tuning for the optimization of the three objectives simultaneously ($\text{TS}_{\text{SP2-SI}}$), we also analyze the use of the multiobjective approach in both rule selection and tuning separately. In practice, we consider chromosomes that are composed of only the C_{S} part for the rule selection (S_{SP2}) and the C_{T} part for the tuning of MFs ($\text{T}_{\text{SP2-SI}}$). Further, their single-objective accuracy-oriented counterparts are also considered in order to analyze the influence of the interpretability criteria in the most difficult one (accuracy).

Clearly, it would make no sense to consider either the GM3M objective when no tuning is performed or the NR objective when no rule selection is performed. It is assumed that the approaches that perform only rule selection have the maximum semantic interpretability and those that perform tuning have the worst NR. Accordingly, the approaches that consider only rule selection should be compared in the accuracy–complexity (MSE–NR) plane, while the approaches that consider only tuning should be compared in the accuracy–semantic (MSE–GM3M) plane. In the case of the proposed method, i.e., $\text{TS}_{\text{SP2-SI}}$, which uses the three objectives, we project the solutions that are obtained in both planes, accuracy–complexity and accuracy–semantic, subsequently removing the dominated solutions that appear from these projections. This way, the methods that perform rule selection and tuning concurrently can be compared with the methods that perform only rule selection in the accuracy–complexity plane and with those that perform only tuning in the accuracy–semantic plane. Some researchers have also used these kinds of

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

TABLE II
METHODS THAT ARE CONSIDERED FOR COMPARISON WITH CLASSICAL TUNING

Method	Ref.	Description	Objectives
WM	[54]	Wang & Mendel Algorithm (Initial RB Generation)	—
		Single-Objective Methods for Post-processing	
S	[35]	Genetic Rule Selection	$MSE_{/2}$
T	[35]	Genetic Tuning of Parameters	$MSE_{/2}$
TS	[35]	Genetic Tuning and Rule Selection	$MSE_{/2}$
		Multi-Objective Evolutionary Algorithms for Post-processing	
$\mathbf{S}_{SP2}$	—	Rule Selection by SPEA2	$MSE_{/2}$ / NR
$\mathbf{T}_{SP2-SI}$	—	Tuning with semantic by SPEA2	$MSE_{/2}$ / GM3M
$\mathbf{TS}_{SP2-SI}$	Proposed here	Tuning and Rule Selection with semantic by SPEA2	$MSE_{/2}$ / NR / GM3M

projections for graphical representation when three objectives are optimized simultaneously [18].

In all the experiments, we adopted a *fivefold cross-validation model*, i.e., we randomly split the dataset into five folds, each containing 20% of the patterns of the dataset, and used four folds for training and one for testing.[1] For each of the possible five different partitions (training/test), the algorithm was applied six times, considering a different seed for the random-number generator. Therefore, we consider the average results of 30 runs. In the case of methods with a multiobjective approach, for each dataset and for each trial, we generate the approximated Pareto front in the corresponding objective planes. Then, we focus on three representative points: the most interpretable (MAX INT), the median (MEDIAN INT/ACC), and the most accurate in training (MAX ACC) points. For each representative point, we compute the mean values over the 30 trials of the MSEs on the training and test sets (i.e., $\text{MSE}^{\text{tra}}_{/2}$ and $\text{MSE}^{\text{tst}}_{/2}$), the NR, and/or the GM3M index, depending on the objective planes that are involved. For the single-objective-based approaches, we compute the same mean values over the 30 solutions that are obtained for each dataset. These three points are representative positions on each plane, i.e., accuracy–complexity or accuracy–semantic, and they have been considered only to perform a statistical analysis on the different planes. Besides, the final user could select the most appropriate solution from the final Pareto front by also looking for a tradeoff between NR and GM3M, depending on its own preferences.

In order to assess whether significant differences exist among the results, we adopt statistical analysis [47]–[50] and, in particular, nonparametric tests, according to the recommendations made in [47] and [48], where a set of simple, safe, and robust nonparametric tests for statistical comparisons of classifiers has been introduced. For pairwise comparison, we use Wilcoxon's signed-ranks test [56], [57], and for multiple comparisons, we will employ different approaches, including Friedman's test [58], Iman and Davenport's test [59], and Holm's method [60]. A detailed description of these tests is presented in Appendix. To perform the tests, we use a level of confidence $\alpha = 0.1$. In particular, Wilcoxon's test is based on computing the differences between two sample means (typically, mean test errors that are obtained by a pair of different algorithms on different datasets). In the classification framework, these differences are well defined since these errors are in the same domain. In our case, to have well-defined differences in $\text{MSE}_{/2}$ and NR (it is not necessary in the case of GM3M), we propose to adopt a normalized difference DIFF, which is defined as

$$\text{DIFF} = \frac{\text{Mean(Other)} - \text{Mean(Reference Algorithm)}}{\text{Mean(Other)}} \quad (9)$$

where $\text{Mean}(x)$ represents either the $\text{MSE}_{/2}$ or the NR means that are obtained by the x algorithm. This difference expresses the improvement in percentage of the reference algorithm.

TABLE III
INITIAL RESULTS OBTAINED BY WM

Dataset	NR	$MSE^{tra}_{/2}$	$MSE^{tst}_{/2}$	σ_{tra}	σ_{tst}
PLA	14.8	3.434	3.557	0.265	0.235
QUA	53.6	0.0258	0.0267	0.001	0.002
ELE	65.0	57606	57934	2841	4733
ABA	68	8.407	8.422	0.443	0.545
STP	122.8	9.074	9.042	0.486	0.809
WAN	156.0	16.063	16.393	0.961	1.700
WIZ	104.8	6.944	7.368	0.720	0.909
MOR	77.6	0.985	0.973	0.129	0.090
TRE	75.0	1.636	1.631	0.121	0.181

The average results of the initial FRBSs, along with their standard deviations (reference results), which are obtained by WM in the five folds, are shown in Table III. In the case of the studied postprocessing algorithms, the values of the input parameters that are considered by the single-objective methods are as follows: population size of 61, 100 000 evaluations, 0.6 as crossover probability, and 0.2 as mutation probability per chromosome. In the case of the MOEAs, these are the following: population size of 200, external population size of 61, 100 000 evaluations, 0.2 as mutation probability, and 30 bits per gene for the Gray codification.

B. Analysis on the Tuning of MFs and the Semantic-Based Index: GM3M

This section analyzes the performance of the methods that perform only tuning of the MFs. Table IV shows the results obtained by T and the results obtained by $\text{T}_{\text{SP2-SI}}$ in the three representative points of the accuracy–semantic plane, which are used further for a statistical analysis of the

[1] The corresponding data partitions (fivefold) for these datasets are available at the KEEL project Web page [55]: http://sci2s.ugr.es/keel/datasets.php.

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
RESULTS OBTAINED BY THE METHODS THAT PERFORM ONLY TUNING OF MFS

Dataset	Method	MAX INT				MEDIAN (INT/ACC)				MAX ACC			
		$MSE^{tra}_{/2}$	$MSE^{tst}_{/2}$	GM3M	(δ γ ρ)	$MSE^{tra}_{/2}$	$MSE^{tst}_{/2}$	GM3M	(δ γ ρ)	$MSE^{tra}_{/2}$	$MSE^{tst}_{/2}$	GM3M	(δ γ ρ)
PLA	T		-				-			1.200	1.251	0.259	(0.11 0.30 0.69)
	T_{SP2-SI}	3.349	3.474	0.985	(0.99 0.98 0.99)	1.572	1.614	0.780	(0.79 0.76 0.80)	1.194	**1.242**	**0.363**	(0.23 0.44 0.72)
QUA	T		-				-			0.0175	**0.0183**	0.097	(0.02 0.12 0.61)
	T_{SP2-SI}	0.0251	0.0260	0.953	(0.96 0.94 0.96)	0.0185	0.0194	0.607	(0.52 0.57 0.75)	0.0175	**0.0183**	**0.109**	(0.01 0.30 0.63)
ELE	T		-				-			17020	21027	0.225	(0.06 0.34 0.69)
	T_{SP2-SI}	56529	56983	0.988	(0.99 0.98 0.99)	24671	26870	0.735	(0.72 0.70 0.79)	15884	**19257**	**0.319**	(0.22 0.40 0.65)
ABA	T		-				-			2.688	2.770	0.144	(0.03 0.18 0.64)
	T_{SP2-SI}	7.886	7.897	0.965	(0.97 0.95 0.97)	3.563	3.611	0.801	(0.83 0.72 0.86)	2.648	**2.744**	**0.298**	(0.18 0.39 0.62)
STP	T		-				-			0.904	1.072	**0.095**	(0.02 0.08 0.60)
	T_{SP2-SI}	8.867	8.834	0.975	(0.98 0.96 0.98)	2.356	2.450	0.706	(0.73 0.65 0.76)	0.797	**0.921**	0.090	(0.01 0.21 0.60)
WAN	T		-				-			1.928	**2.287**	0.144	(0.03 0.19 0.68)
	T_{SP2-SI}	14.683	15.537	0.947	(0.95 0.93 0.95)	3.847	4.650	0.707	(0.69 0.68 0.77)	1.693	2.566	**0.240**	(0.11 0.39 0.68)
WIZ	T		-				-			1.103	1.561	0.155	(0.05 0.14 0.67)
	T_{SP2-SI}	6.516	6.729	0.949	(0.95 0.93 0.96)	1.892	2.037	0.697	(0.65 0.69 0.77)	1.048	**1.243**	**0.260**	(0.10 0.41 0.66)
MOR	T		-				-			0.050	0.061	0.168	(0.04 0.20 0.66)
	T_{SP2-SI}	0.947	0.935	0.975	(0.98 0.97 0.98)	0.245	0.254	0.749	(0.77 0.67 0.82)	0.036	**0.043**	**0.183**	(0.06 0.17 0.72)
TRE	T		-				-			0.052	0.063	**0.152**	(0.04 0.16 0.68)
	T_{SP2-SI}	1.527	1.522	0.951	(0.95 0.94 0.97)	0.303	0.310	0.730	(0.75 0.65 0.82)	0.047	**0.061**	0.113	(0.02 0.20 0.70)

TABLE V
WILCOXON'S TEST: T (R^+) VERSUS $\text{T}_{\text{SP2-SI}}$ (R^-) ON GM3M AND $\text{MSE}^{\text{tst}}_{/2}$ AT MAX ACC

Comparison	Measure	R^+	R^-	Hypothesis ($\alpha = 0.1$)	p-value
T vs. T_{SP2-SI}	GM3M	5	40	Rejected	0.038
T vs. T_{SP2-SI}	$MSE^{tst}_{/2}$	6	39	Rejected	0.051

multiobjective methods. In addition to the semantic-based index, i.e., GM3M, we show the mean values of the three measures that comprise the index, δ, γ, and ρ. [In any event, we should take into account that $(\sum_{i=1}^{30} \text{GM3M}_i/30) \neq \sqrt[3]{(\sum_{i=1}^{30} \delta_i/30)(\sum_{i=1}^{30} \gamma_i/30)(\sum_{i=1}^{30} \rho_i/30)}$.]

Table V shows the results of the Wilcoxon test on the test error and the GM3M measures for T and $\text{T}_{\text{SP2-SI}}$ at MAX ACC. The results show that $\text{T}_{\text{SP2-SI}}$ outperforms T on the test error and GM3M. The null hypothesis that is associated with Wilcoxon's test is rejected ($p < \alpha$) in both cases in favor of $\text{T}_{\text{SP2-SI}}$ due to the differences between R^+ and R^-. This is due to the complex search space that the parametric tuning of MFs involves. The use of both objectives and the modified SPEA2 algorithm helps improve the exploration/exploitation tradeoff to find more optimal solutions.

Fig. 3 shows a representative example in ELE (same data partition and seed) of a DB that is obtained with T and three DBs that are obtained with $\text{T}_{\text{SP2-SI}}$, with the first one with the most interpretable solution, the second one with the median solution, and the last one with the most accurate solution. The DBs obtained are shown in black and the initial DB is shown in gray. To ease graphic representation, the MFs are labeled from "l1" to "l5." Nevertheless, such MFs are associated with a linguistic meaning that is determined by an expert. With these examples, we show the expected correlation between the GM3M index and the semantic interpretability of the obtained DBs. It is quite interesting that the solution with the highest interpretability obtains about a 37% improvement in test with respect to WM and a value of GM3M near 1.

C. Analysis of the Rule Selection

In this section, we present a brief study on the methods that perform only rule selection. Table VI shows the results that are obtained by S and the results that are obtained by S_{SP2} in the three representative points of the accuracy–complexity plane.

In order to assess whether we can conclude that S_{SP2} statistically outperforms S in terms of test error and NR measure, we apply Wilcoxon's test to the results achieved by these algorithms in the most accurate solutions. Table VII shows the results of the application of Wilcoxon's test on these measures. The null hypothesis that is associated with the Wilcoxon's test is now accepted ($p > \alpha$) in both cases. Thus, we can conclude that the results achieved by S and S_{SP2} are statistically different neither on the test error nor on the NR measure. In this case, the search space is well handled by both approaches since equivalent results are obtained by considering the most accurate solutions of the obtained Pareto fronts. In any event, S_{SP2} is able to obtain a set of valid solutions with different accuracy–complexity tradeoffs.

D. Analysis of the Interaction of the Tuning With Rule Selection

This section analyzes the results of the proposed method, i.e., $\text{TS}_{\text{SP2-SI}}$, which performs both rule selection and tuning of the MFs simultaneously, with respect to its single-objective counterpart, i.e., TS. As was explained in Section VI-A, we show the three representative points in Table VIII in the accuracy–semantic and the accuracy–complexity objective planes. This

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

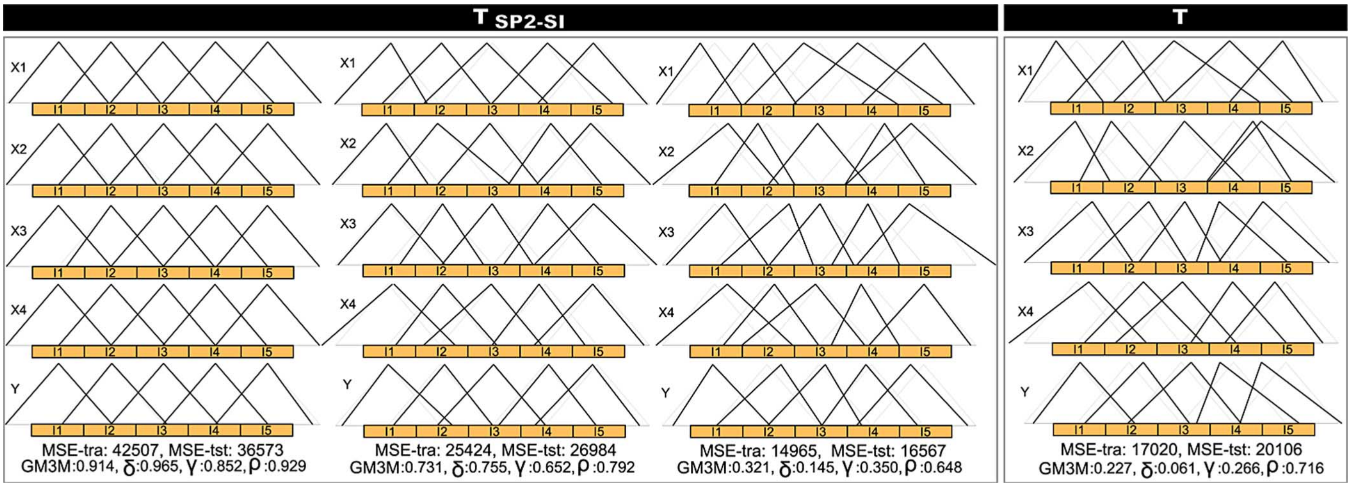
T SP2-SI
T
X1
X2
X3
X4
Y
l1
l2
l3
l4
l5
MSE-tra: 42507, MSE-tst: 36573
GM3M:0.914, δ:0.965, γ:0.852, ρ:0.929
MSE-tra: 25424, MSE-tst: 26984
GM3M:0.731, δ:0.755, γ:0.652, ρ:0.792
MSE-tra: 14965, MSE-tst: 16567
GM3M:0.321, δ:0.145, γ:0.350, ρ:0.648
MSE-tra: 17020, MSE-tst: 20106
GM3M:0.227, δ:0.061, γ:0.266, ρ:0.716

Fig. 3. DB obtained with T and three representative DBs obtained with $\mathrm{T_{SP2\text{-}SI}}$ from one run in ELE.

TABLE VI
RESULTS OBTAINED BY THE METHODS THAT PERFORM ONLY RULE SELECTION

Dataset	Method	MAX INT			MEDIAN (INT/ACC)			MAX ACC		
		NR	$MSE^{tra}_{/2}$	$MSE^{tst}_{/2}$	NR	$MSE^{tra}_{/2}$	$MSE^{tst}_{/2}$	NR	$MSE^{tra}_{/2}$	$MSE^{tst}_{/2}$
PLA	S		-			-		12.6	2.416	**2.416**
	S_{SP2}	6.0	7.042	7.042	9.0	3.473	3.473	**12.0**	2.416	**2.416**
QUA	S		-			-		30.3	0.0220	0.0321
	S_{SP2}	17.4	0.0251	0.0260	24.1	0.0222	0.0234	**28.6**	0.0220	**0.0234**
ELE	S		-			-		40.9	41184	43049
	S_{SP2}	23.7	75486	81764	32.2	44980	46692	**39.0**	41608	**42987**
ABA	S		-			-		**17.7**	4.818	**4.810**
	S_{SP2}	9.9	11.124	11.093	16.4	5.240	5.227	22.2	5.071	5.052
STP	S		-			-		**37.0**	2.532	2.610
	S_{SP2}	30.4	3.741	3.786	45.9	1.521	1.605	62.7	1.446	**1.540**
WAN	S		-			-		47.9	6.418	**7.363**
	S_{SP2}	15.2	12.721	13.564	26.7	6.715	7.966	**38.7**	6.350	7.741
WIZ	S		-			-		44.0	3.036	6.909
	S_{SP2}	14.0	7.333	7.819	24.5	3.261	3.814	**33.7**	3.029	**3.499**
MOR	S		-			-		19.2	0.157	**0.165**
	S_{SP2}	4.2	1.485	1.529	10.1	0.320	0.330	**17.2**	0.252	0.255
TRE	S		-			-		19.7	0.251	**0.257**
	S_{SP2}	4.3	3.091	3.060	11.0	0.396	0.411	**18.5**	0.326	0.342

TABLE VII
WILCOXON'S TEST: S (R^+) VERSUS $\mathrm{S_{SP2}}$ (R^-) ON NR AND $\mathrm{MSE^{tst}_{/2}}$ AT MAX ACC

Comparison	Measure	R^+	R^-	Hypothesis ($\alpha = 0.1$)	p-value
S vs. S_{SP2}	NR	17	28	Accepted	0.515
S vs. S_{SP2}	$MSE^{tst}_{/2}$	23	22	Accepted	0.953

allows further comparisons with the approaches that perform only rule selection and those that perform only tuning. In both cases, the values at the point MAX ACC coincide. The results of the single-objective counterpart algorithm, i.e., TS, are also shown in this table.

This time, we can compare the results from $\mathrm{TS_{SP2\text{-}SI}}$ and TS on the three objective measures. Table IX shows the results of Wilcoxon's test for the most accurate point MAX ACC on them. For each measure, $\mathrm{TS_{SP2\text{-}SI}}$ clearly outperforms TS. The null hypothesis for Wilcoxon's test in all the cases has been rejected in favor of $\mathrm{TS_{SP2\text{-}SI}}$, with a very small p-value, which supports our conclusion with a high degree of confidence. It seems logical that by including NR and GM3M in the multiobjective approach, the interpretability should be better in the obtained FRBSs. However, they are also better in the accuracy objective. The use of the different measures to obtain a set of solutions with different tradeoffs helps maintain a higher diversity that promotes the derivation of more optimal solutions. Therefore, from these results and the results in the previous sections, we can conclude that in the approaches that consider tuning, it is preferable to use a multiobjective approach, including the proposed interpretability measures since we can obtain more interpretable and more accurate FRBSs than those obtained by the single-objective accuracy-oriented counterpart algorithms.

In Fig. 4, we represent some DBs that are obtained with TS and $\mathrm{TS_{SP2\text{-}SI}}$ in ELE and PLA. See Section VI-B for an explanation of these kinds of figures. In both problems, it is clear that at least the DB with the best accuracy from $\mathrm{TS_{SP2\text{-}SI}}$ is preferable to the one that is obtained by TS, but additional highly transparent DBs are also shown in the case of $\mathrm{TS_{SP2\text{-}SI}}$.

E. Global Analysis on the Most Accurate Solutions: MAX ACC

Once the different approaches have been analyzed individually, all of them have to be compared to determine which of them should be preferred. In order to also include the single-objective-based algorithms, the global analysis is performed on the most accurate solutions. Since we will compare more than two algorithms, on this occasion, we use nonparametric tests for multiple comparisons. In order to perform a multiple comparison, it is necessary to check whether any of the results obtained by the algorithms present any inequality. In the case of finding some, we can know, by using a *post hoc* test, which algorithms partners' average results are dissimilar. We will use the results obtained in the evaluation of the three performance measures that have been presented in the previous sections, and we will define a control algorithm as the best performing algorithm (which obtains the lowest value of ranking that is computed through a Friedman test [58]). In order to test whether significant differences exist among all the mean values, we use Iman and Davenport's test [59]. Finally, we use Holm's [60] *post hoc* test to compare the control algorithm with what remains.

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
RESULTS OBTAINED BY THE METHODS THAT PERFORM BOTH RULE SELECTION AND TUNING OF MFS

Dataset	Method	Plane	*MAX* INT							MEDIAN (INT/ACC)							*MAX* ACC						
		$MSE^{tst}_{/2}$,	NR	$MSE^{tra}_{/2}$	$MSE^{tst}_{/2}$	GM3M	(δ	γ	ρ)	NR	$MSE^{tra}_{/2}$	$MSE^{tst}_{/2}$	GM3M	(δ	γ	ρ)	NR	$MSE^{tra}_{/2}$	$MSE^{tst}_{/2}$	GM3M	(δ	γ	ρ)
PLA	TS				-							-					**13.3**	1.255	1.301	0.251	(0.09	0.28	0.74)
	TS_{SP2-SI}	GM3M	12.8	2.624	2.631	0.966	(0.97	0.95	0.98)	13.2	1.600	1.647	0.832	(0.85	0.80	0.84)	13.7	1.170	**1.227**	**0.535**	(0.40	0.59	0.71)
	TS_{SP2-SI}	NR	7.2	1.981	2.022	0.566	(0.46	0.59	0.75)	10.3	1.335	1.412	0.430	(0.25	0.56	0.71)							
QUA	TS				-							-					33.5	0.0173	0.0428	0.182	(0.04	0.27	0.71)
	TS_{SP2-SI}	GM3M	50.5	0.0245	0.0252	0.943	(0.94	0.93	0.96)	34.9	0.0185	0.0192	0.715	(0.69	0.63	0.84)	**27.2**	0.0173	**0.0182**	**0.275**	(0.10	0.34	0.72)
	TS_{SP2-SI}	NR	19.1	0.0176	0.0184	0.372	(0.18	0.43	0.73)	22.4	0.0174	0.0182	0.304	(0.14	0.35	0.73)							
ELE	TS				-							-					41.3	13387	17784	0.335	(0.14	0.39	0.74)
	TS_{SP2-SI}	GM3M	60.2	50703	51959	0.941	(0.94	0.93	0.95)	40.6	28131	31090	0.798	(0.79	0.77	0.84)	**29.3**	11611	**14851**	**0.528**	(0.41	0.58	0.70)
	TS_{SP2-SI}	NR	17.4	34375	38453	0.623	(0.54	0.62	0.73)	21.7	17619	22099	0.544	(0.43	0.55	0.71)							
ABA	TS				-							-					28.4	2.473	2.582	0.021	(0.08	0.36	0.74)
	TS_{SP2-SI}	GM3M	48.1	6.530	6.576	0.958	(0.95	0.96	0.97)	35.8	3.381	3.453	0.826	(0.83	0.77	0.89)	**16.3**	2.386	**2.513**	**0.450**	(0.29	0.54	0.68)
	TS_{SP2-SI}	NR	7.7	3.317	3.423	0.564	(0.47	0.53	0.73)	11.0	2.554	2.670	0.474	(0.35	0.53	0.69)							
STP	TS				-							-					45.6	0.674	1.194	0.276	(0.14	0.25	0.69)
	TS_{SP2-SI}	GM3M	102.0	7.222	7.243	0.964	(0.96	0.95	0.98)	41.3	2.351	2.409	0.808	(0.83	0.75	0.85)	**32.9**	0.642	**0.775**	**0.364**	(0.25	0.40	0.69)
	TS_{SP2-SI}	NR	14.7	2.012	2.194	0.521	(0.41	0.49	0.77)	20.7	0.923	1.064	0.415	(0.28	0.41	0.72)							
WAN	TS				-							-					72.3	1.674	2.373	0.246	(0.08	0.33	0.70)
	TS_{SP2-SI}	GM3M	128.7	12.779	13.704	0.951	(0.94	0.94	0.97)	94.7	5.114	5.760	0.838	(0.86	0.80	0.86)	**39.3**	1.292	**2.016**	**0.456**	(0.31	0.52	0.72)
	TS_{SP2-SI}	NR	19.6	3.315	3.911	0.540	(0.41	0.56	0.74)	26.4	1.685	2.436	0.489	(0.36	0.49	0.72)							
WIZ	TS				-							-					53.5	1.051	2.386	0.349	(0.16	0.45	0.71)
	TS_{SP2-SI}	GM3M	96.5	6.216	6.448	0.962	(0.96	0.96	0.97)	67.7	2.630	2.680	0.833	(0.82	0.82	0.86)	**29.2**	0.921	**1.095**	**0.493**	(0.34	0.57	0.70)
	TS_{SP2-SI}	NR	9.9	2.599	2.963	0.543	(0.42	0.56	0.73)	15.9	1.361	1.522	0.530	(0.40	0.57	0.73)							
MOR	TS				-							-					34.1	0.031	0.037	0.316	(0.14	0.36	0.76)
	TS_{SP2-SI}	GM3M	55.2	0.608	0.634	0.937	(0.93	0.93	0.96)	29.3	0.168	0.180	0.851	(0.85	0.81	0.89)	**15.4**	0.028	**0.034**	**0.541**	(0.44	0.50	0.76)
	TS_{SP2-SI}	NR	5.1	0.202	0.213	0.587	(0.51	0.53	0.76)	8.8	0.065	0.073	0.538	(0.42	0.53	0.75)							
TRE	TS				-							-					29.9	0.050	0.065	0.319	(0.14	0.33	0.76)
	TS_{SP2-SI}	GM3M	57.5	1.141	1.155	0.957	(0.95	0.95	0.97)	32.1	0.310	0.324	0.866	(0.87	0.83	0.90)	**17.7**	0.040	**0.048**	**0.533**	(0.42	0.51	0.74)
	TS_{SP2-SI}	NR	4.9	0.436	0.448	0.583	(0.51	0.54	0.75)	8.9	0.081	0.083	0.531	(0.41	0.50	0.74)							

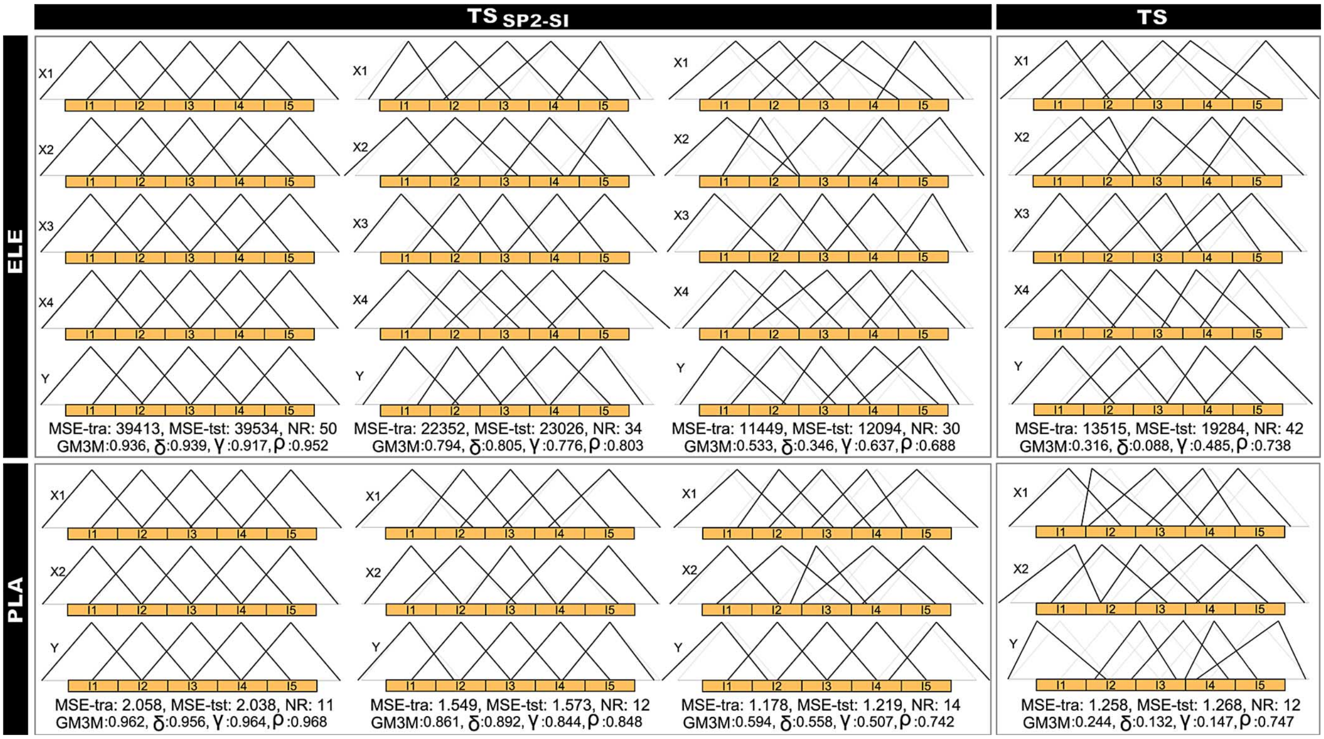
TS SP2-SI
TS
ELE
PLA
X1
X2
X3
X4
Y
l1
l2
l3
l4
l5
MSE-tra: 39413, MSE-tst: 39534, NR: 50
GM3M:0.936, δ:0.939, γ:0.917, ρ:0.952
MSE-tra: 22352, MSE-tst: 23026, NR: 34
GM3M:0.794, δ:0.805, γ:0.776, ρ:0.803
MSE-tra: 11449, MSE-tst: 12094, NR: 30
GM3M:0.533, δ:0.346, γ:0.637, ρ:0.688
MSE-tra: 13515, MSE-tst: 19284, NR: 42
GM3M:0.316, δ:0.088, γ:0.485, ρ:0.738
MSE-tra: 2.058, MSE-tst: 2.038, NR: 11
GM3M:0.962, δ:0.956, γ:0.964, ρ:0.968
MSE-tra: 1.549, MSE-tst: 1.573, NR: 12
GM3M:0.861, δ:0.892, γ:0.844, ρ:0.848
MSE-tra: 1.178, MSE-tst: 1.219, NR: 14
GM3M:0.594, δ:0.558, γ:0.507, ρ:0.742
MSE-tra: 1.258, MSE-tst: 1.268, NR: 12
GM3M:0.244, δ:0.132, γ:0.147, ρ:0.747

Fig. 4. DB obtained with TS and three representative DBs obtained with $\text{TS}_{\text{SP2-SI}}$ from one run in ELE and PLA.

TABLE IX
WILCOXON'S TEST: TS (R^+) VERSUS $\text{TS}_{\text{SP2-SI}}$ (R^-) ON GM3M, NR AND $\text{MSE}^{\text{tst}}_{/2}$ AT MAX ACC

Comparison	Measure	R^+	R^-	Hypothesis ($\alpha = 0.1$)	p-value
TS vs. TS_{SP2-SI}	GM3M	0	45	Rejected	0.008
TS vs. TS_{SP2-SI}	NR	1	44	Rejected	0.011
TS vs. TS_{SP2-SI}	$MSE^{tst}_{/2}$	0	45	Rejected	0.008

TABLE X
RANKINGS OBTAINED THROUGH FRIEDMAN'S TEST FOR THE METHODS THAT PERFORM SELECTION ON $\text{MSE}^{\text{tst}}_{/2}$ AND NR MEASURES

Algorithm	Ranking on $MSE^{tst}_{/2}$	Ranking on NR
TS_{SP2-SI}	1.0000	1.4444
TS	2.2222	3.7778
S_{SP2}	3.4444	2.1111
S	3.3333	2.6667

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

TABLE XI
HOLM TABLE FOR THE METHODS THAT PERFORM SELECTION WITH $\alpha = 0.1$ ON $\mathrm{MSE}_{/2}^{\mathrm{tst}}$ AND NR MEASURES

Holm's post-hoc test on $MSE_{/2}^{tst}$						Holm's post-hoc test on NR					
i	Algorithm	z	p	α/i	Hypothesis	i	Algorithm	z	p	α/i	Hypothesis
3	S_{SP2}	4.02	5.90E-5	0.03	Rejected	3	TS	3.83	1.26E-4	0.03	Rejected
2	S	3.83	1.26E-4	0.05	Rejected	2	S	2.01	0.045	0.05	Rejected
1	TS	2.01	0.05	0.1	Rejected	1	S_{SP2}	1.1	0.27	0.1	Accepted

TABLE XII
RANKINGS OBTAINED THROUGH FRIEDMAN'S TEST FOR THE METHODS THAT PERFORM TUNING ON $\mathrm{MSE}_{/2}^{\mathrm{tst}}$ AND GM3M MEASURES

Algorithm	Ranking on $MSE_{/2}^{tst}$	Ranking on GM3M
TS_{SP2-SI}	1.0000	1.0000
TS	3.2222	2.4444
T_{SP2-SI}	2.5555	3.0
T	3.2222	3.5555

As explained in Section VI-A, the approaches that consider rule selection should be compared in the accuracy–complexity plane, while the approaches that consider tuning should be compared in the accuracy–semantic plane. For this reason, we perform two studies: the first one on the methods that perform rule selection and the second one on the methods that perform tuning. Obviously, TS and the proposed approach, i.e., $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$, are included in both studies by using their projections (see Section VI-A).

1) Analysis of the Methods That Perform Rule Selection—Accuracy–Complexity Plane: Table X shows the rankings of the different methods that are considered in this study. Iman–Davenport's test tells us that significant differences exist among the results observed in all datasets, with p-values (3.990E$-$8) and (8.214E$-$5) on $\mathrm{MSE}_{/2}^{\mathrm{tst}}$ and NR, respectively. The best ranking is obtained by $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$ in both measures: test error and NR.

We now apply Holm's test to compare the best ranking method in each case with the remaining methods. Table XI presents these results, where, the algorithms are ordered with respect to the obtained z-value. Holm's test rejects the hypothesis of equality with the rest of the methods in $\mathrm{MSE}_{/2}^{\mathrm{tst}}$ ($p < \alpha/i$). It also rejects the hypothesis with TS and S in NR. From these results, we can state that $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$ outperforms the remaining methods in both accuracy and complexity, except in the case of $\mathrm{S}_{\mathrm{SP2}}$ that should be considered to be equivalent in terms of NR. However, we can ensure that under these conditions, $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$ dominates $\mathrm{S}_{\mathrm{SP2}}$. It is also interesting to note the ranking position that is obtained by TS on NR. It shows that some unnecessary or inadequate rules cannot be removed by the single-objective approach.

2) Analysis of the Methods That Perform Tuning—Accuracy–Semantic Plane: In this study, Table XII shows the rankings (through Friedman's test) of the four algorithms considered. The p-values computed using Iman–Davenport's test [(8.171E$-$6) and (6.956E$-$7)] imply that there are statistical differences among the results on $\mathrm{MSE}_{/2}^{\mathrm{tst}}$ and GM3M, respectively. $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$ is better in ranking for both measures. In both cases, Holm's test (see Table XIII) rejects the null hypothesis with all the remaining methods. The best method is again $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$, which obtains the best results for these two objectives. Finally, since the proposed approach is the best in both planes, we can conclude that this method is preferable to the remaining approaches to obtain accurate and simple FRBSs, thus maintaining a good level of semantic interpretability.

F. Graphical and Statistical Analysis of the Pareto Fronts

Since we perform 30 trials with different training and test partitions, it would not be readable to show all the Pareto fronts. Thus, to have a glimpse of the trends of the Pareto fronts in the accuracy–complexity and the accuracy–semantic planes, we plot the MAX INT, the MEDIAN (INT/ACC), and the MAX ACC points for each MOEA and for each dataset in Fig. 5. We also show the solutions that are generated by the single-objective methods.

The analysis of Fig. 5 shows that the approximations of the Pareto fronts that are achieved by $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$ are, in general, below the approximations of the Pareto fronts that are obtained by the other MOEAs. To compare in detail the different MOEAs with respect to the MAX INT and MEDIAN (INT/ACC) points, we show the results of the application of the Wilcoxon test on these points in Table XIV for the MOEAs that perform rule selection, i.e., $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$ and $\mathrm{S}_{\mathrm{SP2}}$. We observe a behavior that is very similar to the MAX ACC point, i.e., $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$ outperforms $\mathrm{S}_{\mathrm{SP2}}$ in all the cases, except for NR in the most interpretable point.

With regard to the MOEAs that perform tuning, we show the results of the application of the Wilcoxon test for the same points in Table XV. At the MEDIAN (INT/ACC) point, the null hypothesis that is associated with the Wilcoxon test is rejected ($p < \alpha$) in GM3M, although the results achieved by $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$ and $\mathrm{T}_{\mathrm{SP2\text{-}SI}}$ are statistically equivalent on $\mathrm{MSE}_{/2}^{\mathrm{tst}}$, which is the same as that obtained with MAX INT, but the role between both measures is changed. Under these conditions, we can state that the solutions that are obtained by $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$ dominate, in general, the ones that are obtained by $\mathrm{T}_{\mathrm{SP2\text{-}SI}}$ in practically all the parts of the Pareto fronts.

In order to show the actual behavior of the approximated Pareto fronts provided by each MOEA, we show some representative Pareto fronts (the results of a single trial) on two datasets in Fig. 6. In this figure, we plot the solutions from $\mathrm{TS}_{\mathrm{SP2\text{-}SI}}$ in a 3-D way, and we plot the projections of these solutions on all the possible objective planes along with the corresponding comparison methods. In order to retain all the information, the dominated solutions that are obtained from the projections have not been removed. The symbols and colors similar to those used in Fig. 5 have been used in this case.

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

TABLE XIII
HOLM TABLE FOR THE METHODS THAT PERFORM TUNING WITH $\alpha = 0.1$ ON MSE$^{\mathrm{tst}}_{/2}$ AND GM3M MEASURES

Holm's post-hoc test on $MSE^{tst}_{/2}$						Holm's post-hoc test on GM3M					
i	Algorithm	z	p	α/i	Hypothesis	i	Algorithm	z	p	α/i	Hypothesis
3	T	3.65	2.61E-4	0.03	Rejected	3	T	4.20	2.68E-5	0.03	Rejected
2	TS	3.65	2.61E-4	0.05	Rejected	2	T$_{SP2-SI}$	3.29	0.001	0.05	Rejected
1	T$_{SP2-SI}$	2.56	0.01	0.1	Rejected	1	TS	2.37	0.018	0.1	Rejected

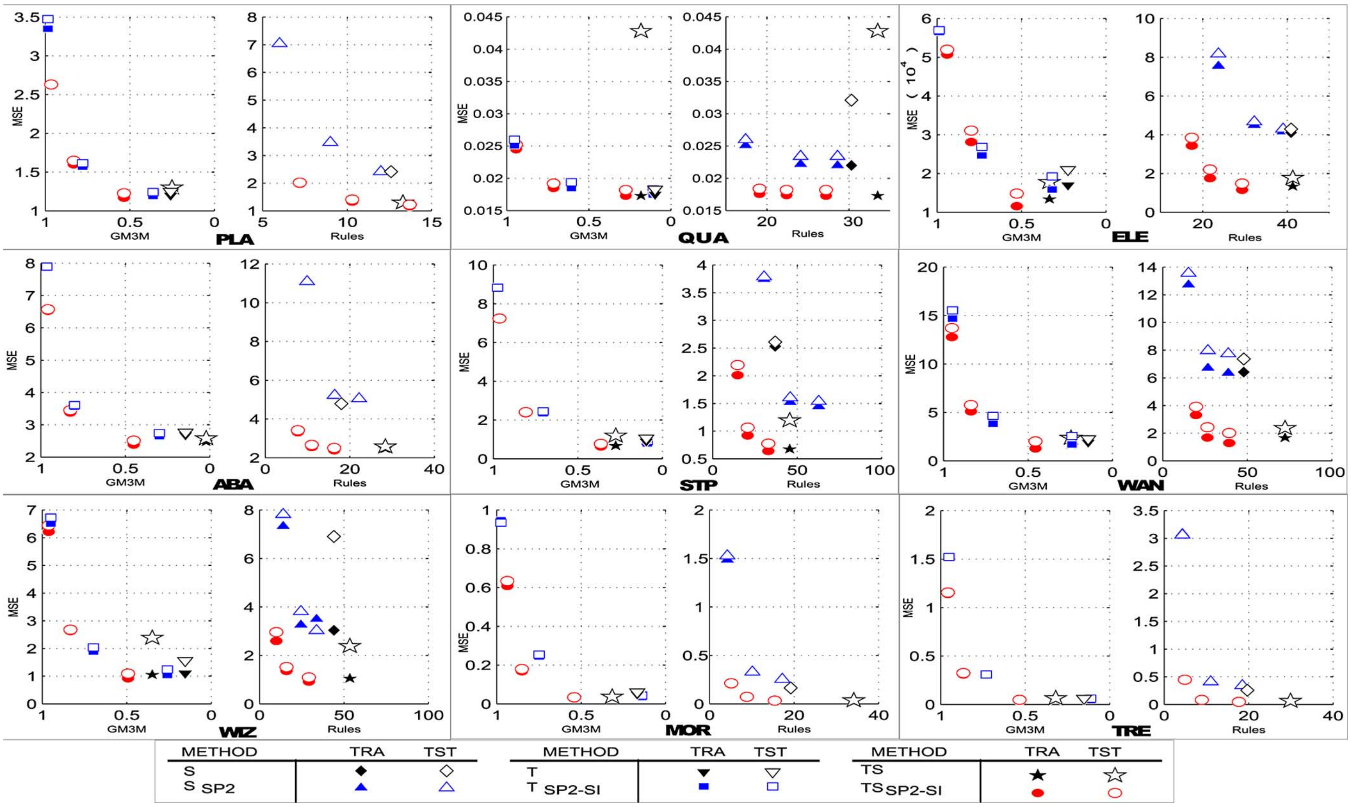

Fig. 5. Averaged Pareto fronts that are obtained in all the problems.

TABLE XIV
WILCOXON'S TEST: S$_{\mathrm{SP2}}$ (R^+) VERSUS TS$_{\mathrm{SP2\text{-}SI}}$ (R^-) ON NR AND MSE$^{\mathrm{tst}}_{/2}$ AT MEDIAN (INT/ACC) AND MAX INT

Comparison	Measure	Pareto solution	R^+	R^-	Hypothesis ($\alpha = 0.1$)	p-value
S_{SP2} vs. TS_{SP2-SI}	NR	MEDIAN (INT/ACC)	3	42	Rejected	0.021
S_{SP2} vs. TS_{SP2-SI}	$MSE^{tst}_{/2}$	MEDIAN (INT/ACC)	1	44	Rejected	0.011
S_{SP2} vs. TS_{SP2-SI}	NR	MAX INT	25	20	Accepted	0.767
S_{SP2} vs. TS_{SP2-SI}	$MSE^{tst}_{/2}$	MAX INT	0	45	Rejected	0.008

TABLE XV
WILCOXON'S TEST: T$_{\mathrm{SP2\text{-}SI}}$ (R^+) VERSUS TS$_{\mathrm{SP2\text{-}SI}}$ (R^-), GM3M AND MSE$^{\mathrm{tst}}_{/2}$ AT MEDIAN (INT/ACC) AND MAX INT

Comparison	Measure	Pareto solution	R^+	R^-	Hypothesis ($\alpha = 0.1$)	p-value
T_{SP2-SI} vs. TS_{SP2-SI}	GM3M	MEDIAN (INT/ACC)	0	45	Rejected	0.008
T_{SP2-SI} vs. TS_{SP2-SI}	$MSE^{tst}_{/2}$	MEDIAN (INT/ACC)	30	15	Accepted	0.374
T_{SP2-SI} vs. TS_{SP2-SI}	GM3M	MAX INT	36	9	Accepted	0.110
T_{SP2-SI} vs. TS_{SP2-SI}	$MSE^{tst}_{/2}$	MAX INT	0	45	Rejected	0.008

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

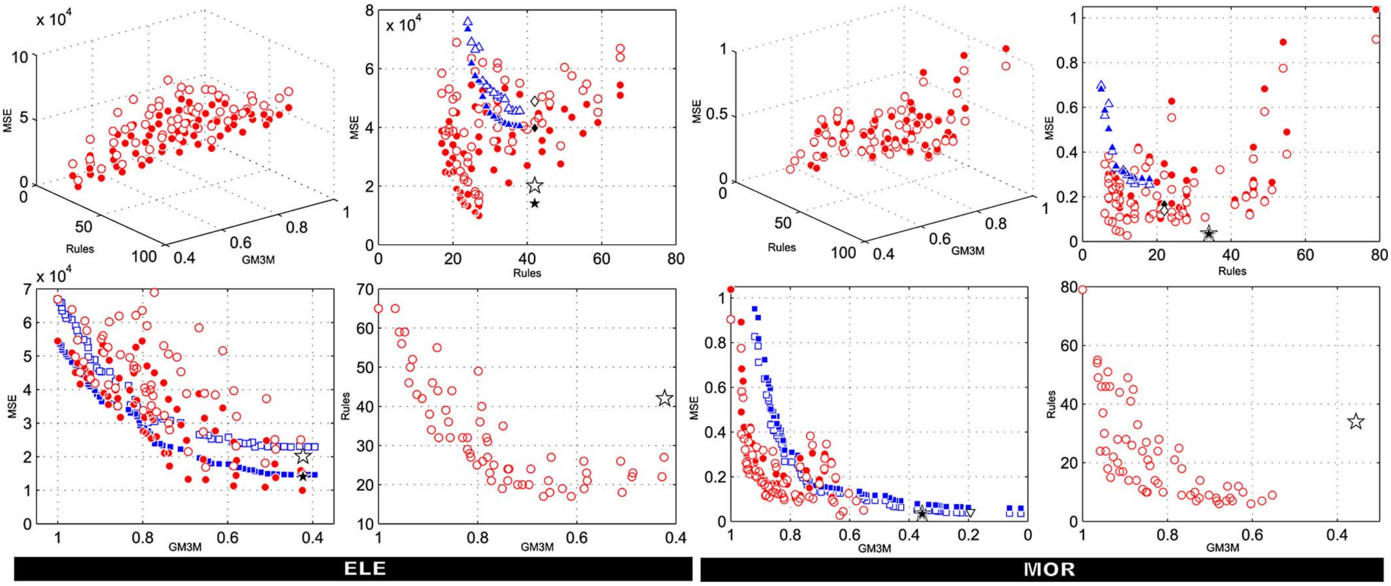

Fig. 6. Example Pareto fronts that are obtained in ELE and MOR problems.

VII. Conclusion

In this paper, we have proposed an index that helps preserve the semantic interpretability of linguistic fuzzy systems, namely, GM3M. The GM3M index is devoted to preserving the original shape of the MFs while a tuning of their definition parameters is performed, and it represents a measure of the quality of the DB. It works on the assumption that the initial DB comprises the appropriate MFs with an associated linguistic meaning (which is usually given by an expert). To this end, we have proposed $\text{TS}_{\text{SP2-SI}}$, which is an effective postprocessing MOEA that is designed to generate a set of FRBSs with different tradeoffs among accuracy, complexity, and semantic interpretability. Three criteria have been considered: the $\text{MSE}_{/2}$, the NR, and the proposed GM3M index. This method performs rule selection and tuning of the MFs simultaneously on a given initial linguistic FRBS.

We have shown that the use of the GM3M index within a multiobjective evolutionary framework helps the tuning approaches obtain more interpretable and, at the same time, more accurate models. Therefore, a multiobjective framework allows us to obtain FRBSs that are characterized by better tradeoffs between accuracy, complexity, and semantic interpretability than the ones that are provided by considering only accuracy as the unique objective.

We should point out that the interaction of rule selection with the tuning of MFs enables the derivation of much more accurate models, while at the same time, the semantic interpretability is maintained to a higher extent. Rule selection allows a major reduction in the system's complexity. Further, we observe that $\text{TS}_{\text{SP2-SI}}$ outperforms all the analyzed methods in all the datasets on the test error, and it achieves better values in GM3M when performing a tuning of the MFs. This way, very interesting solutions have also been obtained with improved accuracy and very high levels of semantic interpretability (near the initial model).

In this sense, this paper has proposed an index to measure the interpretability that is associated with the fuzzy partition along with an RB postprocessing method for obtaining a tradeoff between accuracy and interpretability in linguistic modeling. Working this way follows the final goal pursued by CW by improving the granulation of a continuous variable, which involves a partitioning of the whole into parts, while keeping the meaning of the original words and decreasing the complexity of the RB.

Appendix

On the Use of Nonparametric Tests Based on Rankings

A nonparametric test uses either nominal data, ordinal data, or data represented in an ordinal way of ranking. This does not imply that only they can be used for these types of data. It could be very interesting to transform the data from real values that are contained within an interval to ranking-based data, which is similar to the way a nonparametric test can be applied over typical data of a parametric test when they do not fulfill the necessary conditions that are imposed by the use of the test. In the following, we explain the basic functionality of each nonparametric test used in this study, along with the aim that is pursued by its use.

1) Friedman's test [58]: It is a nonparametric equivalent of the test of repeated-measures analysis of variance (ANOVA). It computes the ranking of the observed results for algorithm (r_j for the algorithm j with k algorithms) for each dataset, assigning the ranking 1 to the best of them and the ranking k to the worst. Under the null hypothesis, which is

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

formed by assuming that the results of the algorithms are equivalent (with similar rankings), Friedman's statistic

$$\mathcal{X}_F^2 = \frac{12N_{ds}}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \qquad (10)$$

is distributed according to $\mathcal{X}_F^2$ with $k-1$ degrees of freedom (DOFs), where $R_j = (1/N_{ds})\sum_i R_i^j$, and N_{ds} is the number of datasets. The critical values for Friedman's statistic coincide with those established in the $\mathcal{X}^2$-distribution when $N_{ds} > 10$ and $k > 5$. On the contrary, the exact values can be seen in [56] and [61].

2) Iman and Davenport's test [59]: It is a metric that is derived from Friedman's statistic given that this last metric produces a conservative undesirable effect. The statistic is

$$\mathcal{F}_F = \frac{(N_{ds}-1)\mathcal{X}_F^2}{N_{ds}(k-1) - \mathcal{X}_F^2} \qquad (11)$$

and it is distributed as an *F*-distribution with $k-1$ and $(k-1)(N_{ds}-1)$ DOFs.

3) Holm's method [60]: This test sequentially checks the hypothesis ordered according to their significance. We will denote the *p*-values ordered by $p_1, p_2, \ldots$ in such a way that $p_1 \leq p_2 \leq \cdots \leq p_{k-1}$. Holm's method compares each p_i with $\alpha/(k-i)$, starting from the most significant *p*-value. If p_1 is less than $\alpha/(k-1)$, the corresponding hypothesis is rejected, and it allows the comparison of p_2 with $\alpha/(k-2)$. If the second hypothesis is rejected, we continue with the process. As soon as a certain hypothesis cannot be rejected, all the remaining hypotheses are maintained as accepted. The statistic for comparing the i algorithm with the j algorithm is

$$z = \frac{(R_i - R_j)}{\sqrt{(k(k+1))/6N_{ds}}}. \qquad (12)$$

The value of z is used to find the corresponding probability from the table of the normal distribution, which is compared with the corresponding value of α.

4) Wilcoxon's signed-rank test: The Wilcoxon signed-rank test is a pairwise test with the aim of detecting significant differences between two sample means: It is analogous to the paired *t*-test in nonparametric statistical procedures. If these means refer to the outputs of two algorithms, then the test practically assesses the reciprocal behavior of the two algorithms [56], [57]. Let d_i be the difference between the performance scores of the two algorithms on the *i*th out of N_{ds} datasets. The differences are ranked according to their absolute values; average ranks are assigned in case of ties. Let R^+ be the sum of ranks for the datasets on which the first algorithm outperformed the second, and let R^- be the sum of ranks for the contrary outcome. Ranks of $d_i = 0$ are split evenly among the sums; if there is an odd number of them, one is ignored:

$$R^+ = \sum_{d_i>0}\text{rank}(d_i) + \frac{1}{2}\sum_{d_i=0}\text{rank}(d_i)$$

$$R^- = \sum_{d_i<0}\text{rank}(d_i) + \frac{1}{2}\sum_{d_i=0}\text{rank}(d_i). \qquad (13)$$

Let T be the smaller of the sums, i.e., $T = \min(R^+, R^-)$. If T is either less than or equal to the value of the distribution of Wilcoxon for N_{ds} DOFs (see [61, Tab. B.12]), the null hypothesis of equality of means is rejected.

REFERENCES

[1] L. A. Zadeh, "From computing with numbers to computing with words—From manipulation of measurements to manipulation of perceptions," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 45, no. 1, pp. 105–119, Jan. 1999.

[2] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *Int. J. Man–Mach. Stud.*, vol. 7, pp. 1–13, 1975.

[3] E. H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic systems," *IEEE Trans. Comput.*, vol. C-26, no. 12, pp. 1182–1191, Dec. 1977.

[4] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning, Part I," *Inf. Sci.*, vol. 8, pp. 199–249, 1975.

[5] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning, Part II," *Inf. Sci.*, vol. 8, pp. 301–357, 1975.

[6] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning, Part III," *Inf. Sci.*, vol. 9, pp. 43–80, 1975.

[7] J. M. Alonso, L. Magdalena, and G. Gonzalez-Rodriguez, "Looking for a good fuzzy system interpretability index. An experimental approach," *Int. J. Approx. Reason.*, vol. 51, no. 1, pp. 115–134, 2009.

[8] J. Casillas, O. Cordón, F. Herrera, and L. Magdalena, Eds., *Accuracy Improvements in Linguistic Fuzzy Modeling* (Studies in Fuzziness and Soft Computing, vol. 129). New York: Springer-Verlag, 2003.

[9] J. Casillas, O. Cordón, F. Herrera, and L. Magdalena, Eds., *Interpretability Issues in Fuzzy Modeling* (Studies in Fuzziness and Soft Computing, vol. 128). New York: Springer-Verlag, 2003.

[10] C. Mencar and A. Fanelli, "Interpretability constraints for fuzzy information granulation," *Inf. Sci.*, vol. 178, no. 24, pp. 4585–4618, 2008.

[11] S. M. Zhou and J. Q. Gan, "Low-level interpretability and high-level interpretability: A unified view of data-driven interpretable fuzzy system modelling," *Fuzzy Sets Syst.*, vol. 159, no. 23, pp. 3091–3131, 2008.

[12] J. M. Alonso, L. Magdalena, and S. Guillaume, "Hilk: A new methodology for designing highly interpretable linguistic knowledge bases using the fuzzy logic formalism," *Int. J. Intell. Syst.*, vol. 23, no. 7, pp. 761–794, 2008.

[13] R. Alcalá, M. J. Gacto, F. Herrera, and J. Alcalá-Fdez, "A multi-objective genetic algorithm for tuning and rule selection to obtain accurate and compact linguistic fuzzy rule-based systems," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 15, no. 5, pp. 539–557, 2007.

[14] M. Cococcioni, P. Ducange, B. Lazzerini, and F. Marcelloni, "A Pareto-based multi-objective evolutionary approach to the identification of Mamdani fuzzy systems," *Soft Comput.*, vol. 11, pp. 1013–1031, 2007.

[15] M. J. Gacto, R. Alcalá, and F. Herrera, "Adaptation and application of multi-objective evolutionary algorithms for rule reduction and parameter tuning of fuzzy rule-based systems," *Soft Comput.*, vol. 13, no. 5, pp. 419–436, 2009.

[16] H. Ishibuchi, K. Nozaki, N. Yamamoto, and H. Tanaka, "Selecting fuzzy if–then rules for classification problems using genetic algorithms," *IEEE Trans. Fuzzy Syst.*, vol. 3, no. 3, pp. 260–270, Aug. 1995.

[17] H. Ishibuchi, T. Murata, and I. B. Türksen, "Single-objective and two-objective genetic algorithms for selecting linguistic rules for pattern classification problems," *Fuzzy Sets Syst.*, vol. 89, no. 2, pp. 135–150, 1997.

[18] H. Ishibuchi, T. Nakashima, and T. Murata, "Three-objective genetics-based machine learning for linguistic rule extraction," *Inf. Sci.*, vol. 136, pp. 109–133, 2001.

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

[19] H. Ishibuchi and T. Yamamoto, "Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining," *Fuzzy Sets Syst.*, vol. 141, no. 1, pp. 59–88, 2004.

[20] H. Ishibuchi and Y. Nojima, "Analysis of interpretability–accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning," *Int. J. Approx. Reason.*, vol. 44, no. 1, pp. 4–31, 2007.

[21] P. Pulkkinen and H. Koivisto, "Fuzzy classifier identification using decision tree and multiobjective evolutionary algorithms," *Int. J. Approx. Reason.*, vol. 48, no. 2, pp. 526–543, 2008.

[22] U. Bodenhofer and P. Bauer, "A formal model of interpretability of linguistic variables," in *Interpretability Issues in Fuzzy Modeling*, J. Casillas, O. Cordón, F. Herrera, and L. Magdalena, Eds. New York: Springer-Verlag, 2003, pp. 524–545.

[23] A. Botta, B. Lazzerini, F. Marcelloni, and D. Stefanescu, "Context adaptation of fuzzy systems through a multiobjective evolutionary approach based on a novel interpretability index," *Soft Comput.*, vol. 13, no. 5, pp. 437–449, 2009.

[24] F. Cheong and R. Lai, "Constraining the optimization of a fuzzy logic controller using an enhanced genetic algorithm," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 30, no. 1, pp. 31–46, Feb. 2000.

[25] J. Espinosa and J. Vandewalle, "Constructing fuzzy models with linguistic integrity from numerical data—AFRELI algorithm," *IEEE Trans. Fuzzy Syst.*, vol. 8, no. 5, pp. 591–600, Oct. 2000.

[26] P. Fazendeiro, J. V. de Oliveira, and W. Pedrycz, "A multiobjective design of a patient and anaesthetist-friendly neuromuscular blockade controller," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 9, pp. 1667–1678, Sep. 2007.

[27] Y. Jin, W. von Seelen, and B. Sendhoff, "On generating fc^3 fuzzy rule systems from data using evolution strategies," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 29, no. 6, pp. 829–845, Dec. 1999.

[28] C. Mencar, G. Castellano, and A. M. Fanelli, "Distinguishability quantification of fuzzy sets," *Inf. Sci.*, vol. 177, pp. 130–149, 2007.

[29] D. Nauck, "Measuring interpretability in rule-based classification systems," in *Proc. 12th IEEE Int. Conf. Fuzzy Syst.,* 2003, vol. 1, no. 2, pp. 196–201.

[30] M. Setnes, R. Babuška, U. Kaymak, and H. R. van Nauta Lemke, "Similarity measures in fuzzy rule base simplification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 28, no. 3, pp. 376–386, Jun. 1998.

[31] J. V. de Oliveira, "Semantic constraints for membership function optimization," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 29, no. 1, pp. 128–138, Jan. 1999.

[32] J. V. de Oliveira, "Towards neuro-linguistic modeling: Constraints for optimization of membership functions," *Fuzzy Sets Syst.*, vol. 106, no. 3, pp. 357–380, 1999.

[33] R. Alcalá, J. Alcalá-Fdez, and F. Herrera, "A proposal for the genetic lateral tuning of linguistic fuzzy systems and its interaction with rule selection," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 4, pp. 616–635, Aug. 2007.

[34] R. Alcalá, J. Alcalá-Fdez, M. J. Gacto, and F. Herrera, "Rule base reduction and genetic tuning of fuzzy systems based on the linguistic 3-tuples representation," *Soft Comput.*, vol. 11, no. 5, pp. 401–419, 2007.

[35] J. Casillas, O. Cordón, M. J. del Jesus, and F. Herrera, "Genetic tuning of fuzzy rule deep structures preserving interpretability and its interaction with fuzzy rule set reduction," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 1, pp. 13–29, Feb. 2005.

[36] O. Cordón and F. Herrera, "A three-stage evolutionary process for learning descriptive and approximate fuzzy logic controller knowledge bases from examples," *Int. J. Approx. Reason.*, vol. 17, no. 4, pp. 369–407, 1997.

[37] H. B. Gürocak, "A genetic-algorithm-based method for tuning fuzzy logic controllers," *Fuzzy Sets Syst.*, vol. 108, no. 1, pp. 39–47, 1999.

[38] F. Herrera, M. Lozano, and J. L. Verdegay, "Tuning fuzzy logic controllers by genetic algorithms," *Int. J. Approx. Reason.*, vol. 12, pp. 299–315, 1995.

[39] C. L. Karr, "Genetic algorithms for fuzzy controllers," *AI Expert*, vol. 6, no. 2, pp. 26–33, 1991.

[40] O. Cordón, F. Herrera, F. Hoffmann, and L. Magdalena, *Genetic Fuzzy Systems. Evolutionary Tuning and Learning of Fuzzy Knowledge Base* (Advances in Fuzzy Systems—Applications and Theory, vol. 19). Singapore: World Scientific, 2001.

[41] C. A. Coello, D. A. V. Veldhuizen, and G. B. Lamont, Eds., *Evolutionary Algorithms for Solving Multi-Objective Problems*. Norwell, MA: Kluwer, 2002.

[42] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*. New York: Wiley, 2001.

[43] F. Herrera, M. Lozano, and J. L. Verdegay, "A learning process for fuzzy control rules using genetic algorithms," *Fuzzy Sets Syst.*, vol. 100, pp. 143–158, 1998.

[44] E. Zitzler, M. Laumanns, and L. Thiele, "SPEA2: Improving the strength Pareto evolutionary algorithm for multiobjective optimization," in *Proc. Evol. Methods Des., Optim. Control Appl. Ind. Probl.*, Barcelona, Spain, 2001, pp. 95–100.

[45] L. J. Eshelman, "The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination," in *Foundations of GAs*, vol. 1, G. Rawlin, Ed., San Mateo, CA: Morgan Kaufman, 1991, pp. 265–283.

[46] K. Deb, A. Pratab, S. Agrawal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[47] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

[48] S. García and F. Herrera, "An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons," *J. Mach. Learn. Res.*, vol. 9, pp. 2677–2694, 2008.

[49] S. García, A. Fernández, J. Luengo, and F. Herrera, "A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability," *Soft Comput.*, vol. 13, no. 10, pp. 959–977, 2009.

[50] S. García, D. Molina, M. Lozano, and F. Herrera, "A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: A case study on the CEC'2005 special session on real parameter optimization," *J. Heuristics*, vol. 15, pp. 617–644, 2009.

[51] A. Krone and H. Taeger, "Data-based fuzzy rule test for fuzzy modelling," *Fuzzy Sets Syst.*, vol. 123, no. 3, pp. 343–358, 2001.

[52] E. H. Ruspini, "A new approach to clustering," *Inf. Control*, vol. 15, no. 1, pp. 22–32, 1969.

[53] L. J. Eshelman and J. D. Schaffer, "Real-coded genetic algorithms and interval-schemata," *Found. Genet. Algorithms*, vol. 2, pp. 187–202, 1993.

[54] L. X. Wang and J. M. Mendel, "Generating fuzzy rules by learning from examples," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, no. 6, pp. 1414–1427, Nov./Dec. 1992.

[55] J. Alcalá-Fdez, L. Sánchez, S. García, M. del Jesus, S. Ventura, J. Garrell, J. Otero, C. Romero, J. Bacardit, V. Rivas, J. Fernández, and F. Herrera, "KEEL: A software tool to assess evolutionary algorithms to data mining problems," *Soft Comput.*, vol. 13, no. 3, pp. 307–318, 2009.

[56] D. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. London, U.K./Boca Raton, FL: Chapman & Hall/CRC, 2003.

[57] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics*, vol. 1, pp. 80–83, 1945.

[58] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Stat. Assoc.*, vol. 32, pp. 675–701, 1937.

[59] R. L. Iman and J. H. Davenport, "Approximations of the critical region of the Friedman statistic," *Commun. Stat. A, Theory Methods*, vol. 9, pp. 571–595, 1980.

[60] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.*, vol. 6, pp. 65–70, 1979.

[61] J. Zar, *Biostatistical Analysis*. Upper Saddle River, NJ: Prentice-Hall, 1999.

María José Gacto received the M.Sc. degree in computer science from the University of Granada, Granada, Spain, in 1999.

She is currently with the Intelligent Systems and Data Mining Research Group, Department of Computer Science, University of Jaén, Jaén, Spain. She has authored or coauthored more than 20 international publications. She has worked on several research projects supported by the Spanish government and the European Union. 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 tradeoff between accuracy and interpretability.

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.

Rafael Alcalá received the M.Sc. and Ph.D. degrees in computer science from the University of Granada, Granada, Spain, in 1998 and 2003, respectively.

From 1998 to 2003, he was with the Department of Computer Science, University of Jaén, Jaén, Spain. He is currently an Assistant Professor with the Department of Computer Science and Artificial Intelligence, University of Granada, where he is also a member of the Soft Computing and Intelligent Information Systems Research Group. He has authored or coauthored more than 60 papers published in international journals, book chapters, and presented at conferences. His research interests include multiobjective genetic algorithms and genetic fuzzy systems, particularly the learning/tuning of fuzzy systems for modeling and control with a good tradeoff between accuracy and interpretability, as well as fuzzy association rules. He has coedited the special issue of *Evolutionary Intelligence* on "Genetic fuzzy systems: New advances." He has worked on several research projects supported by the Spanish government and the European Union. He is currently a member of the editorial/reviewer boards of several journals, including the *International Journal of Computational Intelligence Research*, the *Journal of Advanced Research in Fuzzy and Uncertain Systems*, and the *Journal of Universal Computer Science and Applied Intelligence*.

Dr. Alcalá has coedited the special issue of the IEEE TRANSACTIONS ON FUZZY SYSTEMS on "Genetic fuzzy systems: What's next."

Francisco Herrera received the M.Sc. and Ph.D. degrees in mathematics from the University of Granada, Granada, Spain, in 1988 and 1991, respectively.

He is currently a Professor with the Department of Computer Science and Artificial Intelligence, University of Granada. He has authored or coauthored more than 150 papers published in international journals. He is a coauthor of the book *Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases* (World Scientific, 2001). He has coedited five international books and 20 special issues of international journals on different topics of soft computing. He is an Associate Editor of *Information Sciences*, *Mathware and Soft Computing*, *Advances in Fuzzy Systems*, *Advances in Computational Sciences and Technology*, and the *International Journal of Applied Metaheuristics Computing*. He is also an Area Editor of *Soft Computing* (area of genetic algorithms and genetic fuzzy systems) and is a member of the editorial boards of several journals, including *Fuzzy Sets and Systems*, *Applied Intelligence*, *Knowledge and Information Systems*, *Information Fusion*, *Evolutionary Intelligence*, the *International Journal of Hybrid Intelligent Systems*, and *Memetic Computation*. His current research interests include computing with words and decision making, data mining, data preparation, instance selection, fuzzy-rule-based systems, genetic fuzzy systems, knowledge extraction based on evolutionary algorithms, memetic algorithms, and genetic algorithms.

Prof. Herrera is an Associate Editor of the IEEE TRANSACTIONS ON FUZZY SYSTEMS.

Authorized licensed use limited to: UNIVERSIDAD DE GRANADA. Downloaded on June 18,2010 at 10:46:59 UTC from IEEE Xplore. Restrictions apply.