fuzzy classification rules for multiobjective genetic fuzzy rule selection. Section IV shows an experimental study of this method on a set of 15 well-known datasets. Finally, Section V points out some conclusions.

II. Preliminaries: Fuzzy Rule-Based Classifiers Structure and Inference

Let us assume that we have m training (i.e., labeled) patterns $\vec{x}_p = (x_{p1}, \dots, x_{pn})$, $p = 1, 2, \dots, m$ from M classes in an n-dimensional pattern space where x_{pi} is the attribute value of the pth pattern for the ith attribute ($i = 1, \dots, n$). For the simplicity of explanation, we assume that all the attribute values have already been normalized into real numbers in the unit interval [0, 1]. Thus the pattern space of our classification problem is an n-dimensional unit-hypercube $[0, 1]^n$.

For our n-dimensional pattern classification problem, we use fuzzy rules of the following type:

$$R_q: \text{ If } x_1 \text{ is } A_{q1} \text{ and ... and } x_n \text{ is } A_{qn} \text{ then Class } C_q \text{ with } CF_q, \quad (1)$$

where R_q is the label of the qth fuzzy rule, $\vec{x} = (x_1, \dots, x_n)$ is an n-dimensional pattern vector, A_{qi} is an antecedent fuzzy set ($i = 1, \dots, n$), C_q is a class label, and CF_q is a rule weight. We denote the antecedent fuzzy sets of R_q as a fuzzy vector $\vec{A}_q = (A_{q1}, A_{q2}, \dots, A_{qn})$.

Fourteen fuzzy sets are initially considered in four fuzzy partitions with different granularities. Figure 1 depicts these partitions. In addition to those 14 fuzzy sets, we also use the domain interval [0, 1] itself as an antecedent fuzzy set in order to represent a *don't care* condition.

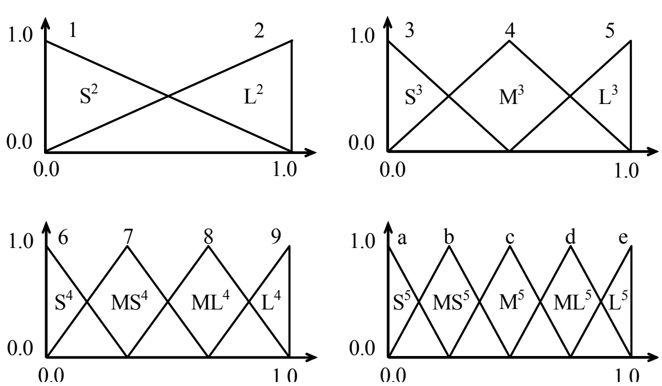

Fig. 1. The fourteen antecedent fuzzy sets considered.

Let S be a set of fuzzy rules of the form in (1). When an input pattern $\vec{x}_p$ is to be classified by S, first we calculate the compatibility grade of $\vec{x}_p$ with the antecedent part $\vec{A}_q = (A_{q1}, A_{q2}, \dots, A_{qn})$ of each fuzzy rule R_q in S using the product operation as,

$$\mu_{\vec{A}_q}(\vec{x}_p) = \mu_{A_{q1}}(x_{p1}) \cdot \ldots \cdot \mu_{A_{qn}}(x_{pn}), \quad (2)$$

where $\mu_{A_{qi}}(\cdot)$ is the membership function of the antecedent fuzzy set A_{qi}. Then a single winner rule R_w is identified using the compatibility grade and the rule weight of each fuzzy rule as

$$\mu_{\vec{A}_w}(\vec{x}_p) \cdot CF_w = \max\{\mu_{\vec{A}_q}(\vec{x}_p) \cdot CF_q \mid R_q \in S\}. \quad (3)$$

The input pattern $\vec{x}_p$ is classified as the consequent class C_w of the winner rule R_w. When multiple fuzzy rules with different consequent classes have the same maximum value in (3), the classification of $\vec{x}_p$ is rejected. If there is no compatible fuzzy rule with $\vec{x}_p$, its classification is also rejected.

III. An Algorithm for Generating Single Granularity-Based Fuzzy Classification Rules

As we have already explained, multiobjective genetic fuzzy rule selection has been based on a previously fixed granularity [10], [11] (five linguistic terms in all the attributes) or multiple granularities [12]. Based on this last approach [12], in this section we propose a mechanism to generate single granularity-based fuzzy classification rules, a nearer to the interpretability approach. The proposed procedure is as follows:

- *Step 1:* Rule extraction with multiple granularities.
- *Step 2:* Specification of single granularity for each attribute based on the extracted rules.
- *Step 3:* Rule extraction with selected single granularities.
- *Step 4:* Multiobjective genetic fuzzy rule selection.

The original multiple granularities based procedure [12] is composed of Steps 1 and 4. Steps 2 and 3 are additional procedures. In Step 1, we extract a fixed short number of rules for each class based on well-known data mining rule evaluation measures [2] and multiple granularities. In Step 2, we select a single granularity for each attribute based on the extracted rules. Then, we extract the final set of candidate rules for each class by using the selected single granularities in Step 3. Step 4 is the same as the original one to perform multiobjective genetic fuzzy rule selection. The next subsections present detailed explanations of these steps.

A. Rule Extraction with Multiple Granularities (Step 1)

Since 14 antecedent fuzzy sets in Figure 1 and an additional *don't care* fuzzy set [0, 1] are used for each attribute of the n-dimensional classification problem, the total number of possible fuzzy rules is 15^n. Among these possible rules, we examine only short fuzzy rules with a small number of antecedent conditions (i.e., short fuzzy rules with many *don't care* conditions) to generate an initial set of candidate rules. In this work, we specify the maximum number of antecedent conditions as three for datasets with less than 30 attributes and two for datasets with more than or equal to 30 attributes.

The consequent class C_q and the rule weight CF_q of each fuzzy rule R_q are specified from training patterns compatible with its antecedent part $\vec{A}_q = (A_{q1}, A_{q2}, \dots, A_{qn})$ in the following heuristic manner [13]. First the confidence of each class for the antecedent part $\vec{A}_q$ is calculated as:

$$c(\mathbf{A}_q \Rightarrow \text{Class } h) = \frac{\displaystyle\sum_{\mathbf{x}_p \in \text{ Class } h} \mu_{\mathbf{A}_q}(\mathbf{x}_p)}{\displaystyle\sum_{p=1}^{m} \mu_{\mathbf{A}_q}(\mathbf{x}_p)}, \quad h = 1, \ldots, M. \tag{4}$$

It should be noted that "$\mathbf{A}_q \Rightarrow$ Class h" means the fuzzy rule with the antecedent part $\vec{A}_q$ and the consequent class h. Then the consequent class C_q is specified by identifying the class with the maximum confidence:

$$c(\mathbf{A}_q \Rightarrow \text{Class } C_q) = \max_{h=1,\,2,\,\ldots,\,M} \{c(\mathbf{A}_q \Rightarrow \text{Class } h)\}. \tag{5}$$

In this manner, we generate the fuzzy rule R_q (i.e., $\mathbf{A}_q \Rightarrow$ Class C_q) with the antecedent part $\vec{A}_q$ and the consequent class C_q. We do not generate any fuzzy rules with the antecedent part $\vec{A}_q$ if there is no compatible training pattern with $\vec{A}_q$.

The rule weight CF_q of each fuzzy rule R_q has a large effect on the performance of fuzzy rule-based classifiers. We use the following specification of CF_q because good results were reported in the literature [14]:

$$CF_q = c(\mathbf{A}_q \Rightarrow \text{Class}\, C_q) - \sum_{\substack{h=1 \\ h \neq C_q}}^{M} c(\mathbf{A}_q \Rightarrow \text{Class}\, h). \tag{6}$$

We do not use the fuzzy rule R_q as a candidate rule if the rule weight CF_q is not positive (i.e., if its confidence is not larger than 0.5).

In the above-mentioned heuristic manner, we can generate a large number of short fuzzy rules as candidate rules in multiobjective fuzzy rule selection (some of them with not interesting properties). In order to directly focus on the most interesting rules, a prescreening procedure is applied to decrease the number of candidate rules. This prescreening procedure is based on well-known rule evaluation measures in the field of data mining [2]: *support* and *confidence*.

For prescreening candidate rules, we use two threshold values: the minimum support and the minimum confidence. We exclude fuzzy rules that do not satisfy these two threshold values. Among short fuzzy rules satisfying these two threshold values, we choose a prespecified number of candidate rules for each class. As a rule evaluation criterion, we use the product of the support $s(R_q)$ and the confidence $c(R_q)$. That is, we choose a prespecified number of the best candidate rules for each class with respect to product $p(R_q) = s(R_q) \cdot c(R_q)$.

B. Single Granularity Specification and Rule Extraction (Steps 2 and 3)

Once a set of candidate rules is obtained based on multiple granularities (Step 1), the original approach [12] goes to Step 4 in order to apply multiobjective fuzzy rule selection. However, there is useful information in the extracted rules that could be used to specify an appropriate single granularity for each attribute. Frequency of the employed granularities in the extracted rules (weighted by the corresponding rule importance) can be used to fix the most promising granularities. For each attribute i ($i = 1, \ldots, n$), we specify the granularity with the highest sum of importance of the rules considering such granularity in the corresponding attribute:

$$Gr(i) = \underset{g=2,\,\ldots,\,5}{\operatorname{argmax}} \left\{ \sum_{Gran(\mathbf{A}_{qi})=g} \operatorname{Imp}(\mathbf{R}_q) \right\}, \tag{7}$$

where $Gran(\mathbf{A}_{qi})$ is the granularity of the partition containing the fuzzy set used in attribute i of rule R_q and $\operatorname{Imp}(R_q)$ is a criterion associated to the importance of the rule in the sum. Many criteria can be considered involving different specification mechanisms:

- Frequency: $\operatorname{Imp}(R_q) = 1, \forall q$.
- Confidence: $\operatorname{Imp}(R_q) = c(R_q), \forall q$.
- Weight: $\operatorname{Imp}(R_q) = CF_q, \forall q$.
- Support: $\operatorname{Imp}(R_q) = s(R_q), \forall q$.
- Product: $\operatorname{Imp}(R_q) = p(R_q), \forall q$.

However, the first three criteria are not recommended since they usually provoke overfitting. We will study the last two criteria as a way to extract more general rules instead of very specific ones, which helps to the generalization ability. In the same way, in order to preferably take into account more general rules we examine two approaches named, 1-ALL approach and 1-2-3 approach, with the two basic criteria (i.e., product and support). Both approaches give priority to granularities in the rules with a single condition, i.e., Equation (7) is applied by only considering size one rules if possible. The difference is only when there is no rule with a single condition in the corresponding attribute. Let us consider the product criterion and the next six rules, where g^i represents any fuzzy set of a partition with granularity i,

R_1 : If x_1 is g^2 and x_2 is g^4 and x_3 is g^3 then Class 1, $p(R_1) : 0.4$.
R_2 : If x_1 is g^4 then Class 2, $p(R_2) : 0.8$.
R_3 : If x_2 is g^3 then Class 2, $p(R_3) : 0.3$.
R_4 : If x_2 is g^2 then Class 1, $p(R_4) : 0.8$.
R_5 : If x_2 is g^3 and x_3 is g^4 then Class 1, $p(R_5) : 0.6$.
R_6 : If x_1 is g^2 and x_2 is g^2 and x_3 is g^3 then Class 1, $p(R_6) : 0.3$.

When we specify a granularity for the first attribute, we first check rule(s) with a single condition related to the first attribute by both approaches (1-ALL and 1-2-3). Since rule R_2 is the only rule in this situation, we select granularity 4 for the first attribute. Next, in the same manner, we can find two rules: R_3 and R_4 for the second attribute. We select Granularity 2 for the second attribute because of the high product value by both approaches. Finally, we select a granularity for the third attribute but there is no rule with a single condition. In 1-ALL approach, we specify a single granularity from all the rules including the third attribute