

Toward Quantitative Definition of Explanation Ability of Fuzzy Rule-based Classifiers

Hisao Ishibuchi

Department of Computer Science and Intelligent Systems
Osaka Prefecture University
1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531, Japan
hisaoi@cs.osakafu-u.ac.jp

Yusuke Nojima

Department of Computer Science and Intelligent Systems
Osaka Prefecture University
1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531, Japan
nojima@cs.osakafu-u.ac.jp

Abstract—Explanation ability of a fuzzy rule-based classifier is its ability to explain why an input pattern is classified as a particular class in a convincing way. This ability is important especially when fuzzy rule-based classifiers are used as support systems for human users. This is because human users often want to know why the current input pattern is classified as a particular class. The explanation ability looks similar to the interpretability. They are, however, clearly different concepts. Whereas the explanation ability is directly related to the classification of each pattern, the interpretability is usually independent of classification results. The interpretability has been taken into account in multiobjective design of fuzzy rule-based classifiers. However, the explanation ability has not been used for fuzzy rule-based classifier design. This is because its quantitative definition is very difficult. In this paper, we discuss various factors that are related to quantitative definition of the explanation ability of fuzzy rule-based classifiers. Using simple numerical examples, we explain that the complexity minimization of fuzzy rule-based classifiers does not always lead to the explanation ability maximization. We also explain that the accuracy of fuzzy rules is related to the explanation ability.

Keywords—pattern classification; fuzzy systems; fuzzy rule-based classifiers; explanation ability; interpretability; complexity.

I. INTRODUCTION

Fuzzy rule-based classifier design has been often handled as the search for accurate and interpretable fuzzy rule-based classifiers. That is, accuracy and interpretability have been used as design criteria in fuzzy rule-based classifier design. These two criteria are combined into a single overall criterion in single-objective approaches [1]-[4] whereas they are handled as different objectives in multiobjective approaches [5]-[9]. Since there exists a tradeoff relation between these two criteria [10], multiobjective approaches usually find a large number of non-dominated fuzzy rule-based classifiers with different tradeoffs (i.e., from simple classifiers with high interpretability and low accuracy to complicated classifiers with high accuracy and low interpretability). It is implicitly assumed that a single fuzzy rule-based classifier is chosen from a large number of obtained non-dominated ones by human users based on their preference in multiobjective approaches. In single-objective approaches, parameter values in a single overall design criterion are often assumed to be specified by human users' preference.

In some early studies on multiobjective fuzzy rule-based classifier design [5], [6], interpretability maximization was

formulated as complexity minimization using complexity measures such as the number of fuzzy rules and the total number of antecedent conditions (i.e., total rule length). Recently more sophisticated interpretability measures have been used in multiobjective approaches to fuzzy rule-based classifier design [7]-[9]. Definition of interpretability itself is currently a hot research issue in fuzzy systems [11]-[19].

Whereas the interpretability of fuzzy rule-based classifiers has been frequently discussed in the literature [11]-[19], the explanation ability has not been discussed in many studies [20]-[24]. In this paper, we use the term “*explanation ability*” to mean the ability of a fuzzy rule-based classifier to explain why an input pattern is classified as a particular class in a convincing way. It should be noted that the interpretability and the explanation ability are different concepts. On the one hand, the interpretability of a fuzzy rule-based classifier is often defined by its structural features (e.g., the shape of membership functions, the overlap of adjacent membership functions, the number of fuzzy rules, and the number of antecedent conditions) independent of the classification of each pattern. On the other hand, the explanation ability is directly related to classification results from its definition: the ability of a fuzzy rule-based classifier to explain why an input pattern is classified as a particular class in a convincing way.

In this paper, we discuss various factors that are related to quantitative definition of the explanation ability of a fuzzy rule-based classifier. We assume the use of a single winner-based fuzzy reasoning method for pattern classification. This means that each input pattern is classified by a single winner rule. Thus the explanation of the classification result of an input pattern is performed by showing the corresponding single winner rule. In this situation, the following factors seem to be related to the explanation ability:

- (1) Compatibility of the input pattern with its winner rule: If the winner rule has high compatibility with the input pattern, it may have high explanation ability since its antecedent conditions can be viewed as describing the input pattern very well (i.e., without large inconsistency).
- (2) Interpretability of the winner rule: If the winner rule has high interpretability, it may have high explanation ability since the explanation is easy for human users to understand.
- (3) Accuracy of the winner rule: If the winner rule has high accuracy, it may have high explanation ability since the

classification result is highly reliable.

- (4) Coverage of the winner rule: If the winner rule classifies many patterns, it may have high explanation ability since the same explanation is applicable to many patterns.

For discussing these factors, this paper is organized as follows. First we explain our fuzzy rule-based classifier with the single winner-based fuzzy reasoning method in Section II. Then we discuss mathematical formulations of the above-mentioned four factors in details in Section III using simple numerical examples. Finally we conclude this paper in Section IV where we also suggest some future research directions.

II. FUZZY RULE-BASED CLASSIFIERS

We assume that we have an M -class pattern classification problem in an n -dimensional normalized hyper-cube $[0, 1]^n$. For this problem, we use fuzzy rules of the following type:

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

where R_q is the label of the q th fuzzy rule, $\mathbf{x}=(x_1, \dots, x_n)$ is an n -dimensional pattern vector, A_{qi} is an antecedent fuzzy set ($i=1,2,\dots,n$), C_q is a class label, and CF_q is a real number in the unit interval $[0, 1]$ which represents a rule weight. The rule weight can be specified by a heuristic manner [25]. It can be also adjusted by a learning algorithm [26]. We use the n -dimensional fuzzy vector $\mathbf{A}_q=(A_{q1}, \dots, A_{qn})$ for representing the antecedent part of the fuzzy rule R_q in (1) in a concise manner.

Let S be a set of fuzzy rules of the type in (1). That is, S is a fuzzy rule-based classifier. When an n -dimensional input pattern $\mathbf{x}_p=(x_{p1}, \dots, x_{pn})$ is presented to S , first the compatibility grade of $\mathbf{x}_p$ with the antecedent part $\mathbf{A}_q$ of each fuzzy rule R_q in S is calculated by the product operator as

$$\mu_{\mathbf{A}_q}(\mathbf{x}_p) = \mu_{A_{q1}}(x_{p1}) \times \dots \times \mu_{A_{qn}}(x_{pn}) \text{ for } R_q \in S, \quad (2)$$

where $\mu_{A_{qi}}(\cdot)$ shows the membership function of A_{qi} . Then a single winner rule $R_{w(\mathbf{x}_p)}$ is identified for $\mathbf{x}_p$ as follows:

$$CF_{w(\mathbf{x}_p)} \times \mu_{\mathbf{A}_{w(\mathbf{x}_p)}}(\mathbf{x}_p) = \max\{CF_q \times \mu_{\mathbf{A}_q}(\mathbf{x}_p) \mid R_q \in S\}, \quad (3)$$

where $w(\mathbf{x}_p)$ denotes the rule index of the winner rule for $\mathbf{x}_p$.

The pattern $\mathbf{x}_p$ is classified by the single winner rule $R_{w(\mathbf{x}_p)}$ as its consequent class. If there is no fuzzy rule with a positive compatibility grade with $\mathbf{x}_p$ (i.e., if $\mathbf{x}_p$ is not covered by any fuzzy rules in S), the classification of $\mathbf{x}_p$ is rejected. The classification of $\mathbf{x}_p$ is also rejected if multiple fuzzy rules with different consequent classes have the same maximum value in the right-hand side of (3). In this case, $\mathbf{x}_p$ is exactly on the classification boundary between the different classes.

We use the single winner-based fuzzy reasoning method in (3) for pattern classification. This is because a responsible fuzzy rule can be always identified for the classification result of each input pattern when we use the single winner-based fuzzy reasoning method. For other fuzzy reasoning methods (e.g., a voting-based method), see [27]-[29] where other types

of fuzzy rules (e.g., fuzzy rules with no rule weights, fuzzy rules with multiple consequent classes) were also explained.

III. EXPLANATION ABILITY

In this section, we discuss each of the four factors related to the explanation ability of a fuzzy rule-based classifier in detail: compatibility, interpretability, accuracy and coverage.

A. A Simple Test Problem and Fuzzy Rule-Based Classifiers

For visual explanation, let us consider a simple three-class two-dimensional pattern classification problem with 20 patterns from each class in Fig. 1. Using three linguistic values (*small*, *medium* and *large*) in Fig. 1 as antecedent fuzzy sets for each of the two attributes, we can generate 3×3 fuzzy rules. Let S_1 be the fuzzy rule-based classifier with those nine fuzzy rules (for simplicity, the rule weight is specified as 1.0 in S_1):

S_1 : Fuzzy rule-based classifier with nine fuzzy rules

- R_1 : If x_1 is *small* and x_2 is *small* then Class 2 with 1.0,
- R_2 : If x_1 is *small* and x_2 is *medium* then Class 2 with 1.0,
- R_3 : If x_1 is *small* and x_2 is *large* then Class 1 with 1.0,
- R_4 : If x_1 is *medium* and x_2 is *small* then Class 2 with 1.0,
- R_5 : If x_1 is *medium* and x_2 is *medium* then Class 2 with 1.0,
- R_6 : If x_1 is *medium* and x_2 is *large* then Class 1 with 1.0,
- R_7 : If x_1 is *large* and x_2 is *small* then Class 3 with 1.0,
- R_8 : If x_1 is *large* and x_2 is *medium* then Class 3 with 1.0,
- R_9 : If x_1 is *large* and x_2 is *large* then Class 3 with 1.0.

The location of each rule is shown in Fig. 2 (a). All the 60 patterns in Fig. 1 are correctly classified by S_1 with these nine rules in S_1 as shown in Fig. 3 (a).

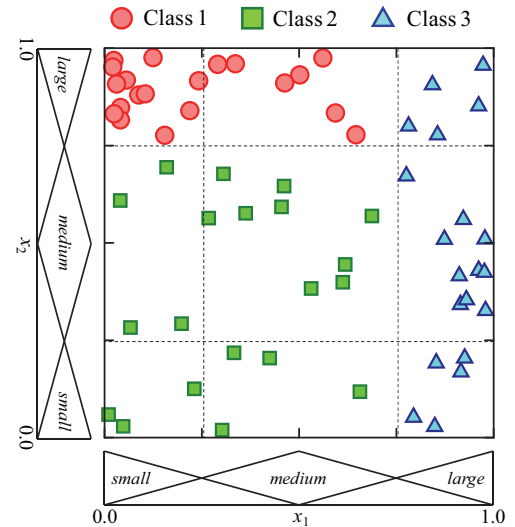


Fig. 1. A simple three-class two-dimensional test problem.

Let us try to simplify the fuzzy rule-based classifier S_1 . We can see from Fig. 1 that all training patterns with large values of x_1 are from Class 3. Accordingly, the last three fuzzy rules R_7 , R_8 and R_9 in S_1 with the same antecedent condition “ x_1 is *large*” have the same consequent class (i.e., Class 3). This

observation suggests us to combine the last three rules R_7 , R_8 , and R_9 of S_1 into a single fuzzy rule:

R_{789} : If x_1 is *large* then Class 3 with 1.0.

Using this fuzzy rule, S_1 can be simplified as follows:

S_2 : Fuzzy rule-based classifier with seven fuzzy rules

- R_1 : If x_1 is *small* and x_2 is *small* then Class 2 with 1.0,
- R_2 : If x_1 is *small* and x_2 is *medium* then Class 2 with 1.0,
- R_3 : If x_1 is *small* and x_2 is *large* then Class 1 with 1.0,
- R_4 : If x_1 is *medium* and x_2 is *small* then Class 2 with 1.0,
- R_5 : If x_1 is *medium* and x_2 is *medium* then Class 2 with 1.0,
- R_6 : If x_1 is *medium* and x_2 is *large* then Class 1 with 1.0,
- R_{789} : If x_1 is *large* then Class 3 with 1.0.

This fuzzy rule-based classifier S_2 is illustrated in Fig. 2 (b). All the 60 patterns in Fig. 1 are correctly classified by S_2 as shown in Fig. 3 (b). From Fig. 3 (a) and Fig. 3 (b), we can see that S_1 and S_2 have slightly different classification boundaries.

If we use the negation of *large* (i.e., “not large”) in the antecedent part, the number of fuzzy rules in S_2 can be further decreased as follows:

S_3 : Fuzzy rule-based classifier with three fuzzy rules

- R_{1245} : If x_1 is *not large* and x_2 is *not large* then Class 2 with 1.0,
- R_{36} : If x_1 is *not large* and x_2 is *large* then Class 1 with 1.0,
- R_{789} : If x_1 is *large* then Class 3 with 1.0.

The membership function of “not large” is defined as

$$\mu_{\text{not large}}(x) = 1 - \mu_{\text{large}}(x) \text{ for } 0 \leq x \leq 1, \quad (4)$$

where $\mu_{\text{large}}(x)$ is the membership function of the antecedent fuzzy set *large*.

The fuzzy rule-based classifier S_3 with the three fuzzy rules is illustrated in Fig. 2 (c). All the 60 patterns are correctly classified by S_3 as shown in Fig. 3 (c). From Fig. 3 (a)-(c), we can see that the three fuzzy rule-based classifiers S_1 , S_2 and S_3 have slightly different classification boundaries.

The fuzzy rule-based classifier S_3 can be viewed as having the following tree-structure:

- If x_1 is *large* then Class 3 with 1.0,
- else if x_2 is *large* then Class 1 with 1.0,
- else Class 2 with 1.0. (5)

A simple fuzzy rule-based classifier can be also obtained by selecting fuzzy rules from S_1 rather than combining fuzzy rules. For example, the following fuzzy rule-based classifier with four fuzzy rules has the same classification boundaries as S_1 with the nine fuzzy rules (see Fig. 2 (d) and Fig. 3 (d)):

S_4 : Fuzzy rule-based classifier with four fuzzy rules

- R_5 : If x_1 is *medium* and x_2 is *medium* then Class 2 with 1.0,
- R_6 : If x_1 is *medium* and x_2 is *large* then Class 1 with 1.0,
- R_8 : If x_1 is *large* and x_2 is *medium* then Class 3 with 1.0,
- R_9 : If x_1 is *large* and x_2 is *large* then Class 3 with 1.0.

Strictly speaking, patterns with $x_1 = 0.0$ or $x_2 = 0.0$ cannot be classified by S_4 whereas those patterns can be classified by S_1 . Thus the classification results by S_1 and S_4 are not exactly the same. However, except for those special cases, S_1 and S_4 have the same classification results as shown in Fig. 3.

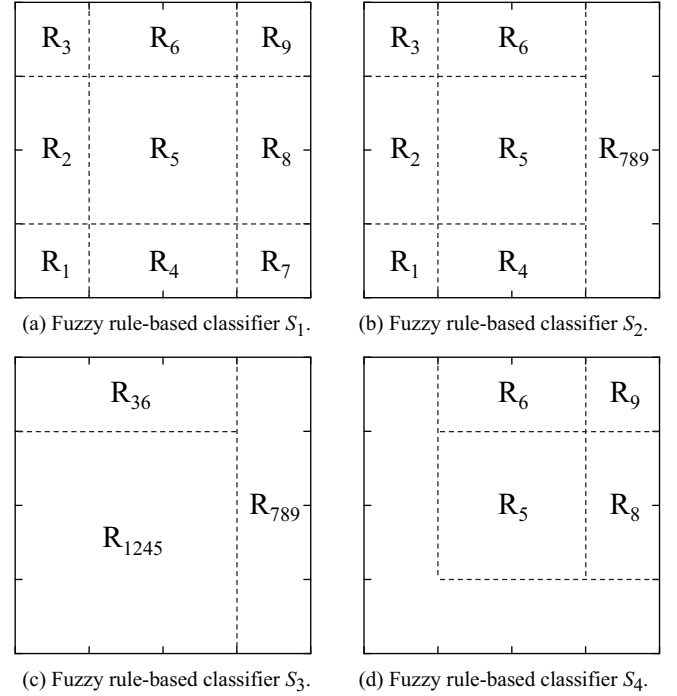


Fig. 2. Fuzzy rules in S_1 , S_2 , S_3 and S_4 .

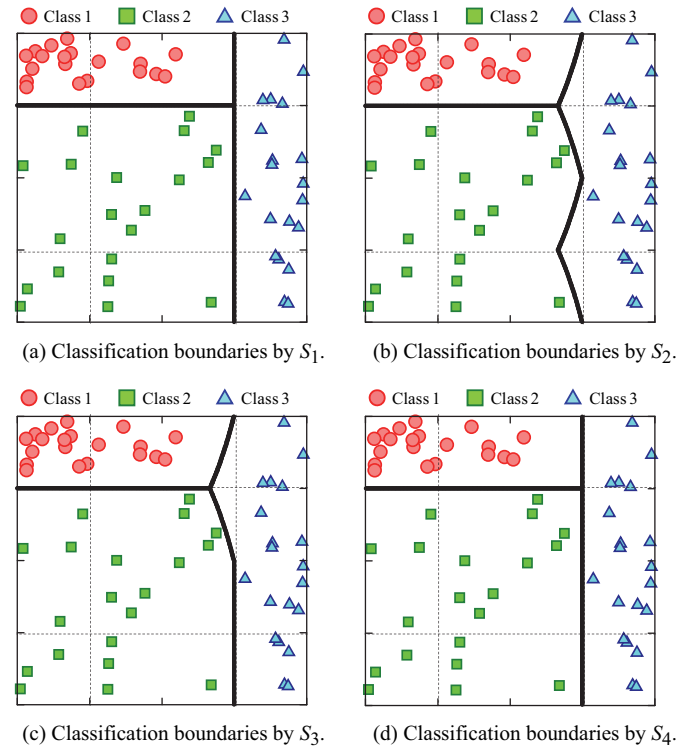


Fig. 3. Classification boundaries by S_1 , S_2 , S_3 and S_4 .

Since S_4 is a subset of S_1 , it is clear that S_4 is simpler (i.e., less complicated) than S_1 (see Fig. 2 (a) and Fig. 2 (d)). It is also clear that S_2 is simpler than S_1 since the three fuzzy rules R_7 , R_8 and R_9 in S_1 are merged into the single fuzzy rule R_{789} in S_2 . However, the comparison of S_3 with the other fuzzy rule-based classifiers is not easy since S_3 has the special antecedent fuzzy set “*not large*”. If we use the number of fuzzy rules as a complexity measure, we have the following order of the four fuzzy rule-based classifiers with respect to their complexity:

$$\text{Complex}(S_1) > \text{Complex}(S_2) > \text{Complex}(S_4) > \text{Complex}(S_3), \quad (6)$$

where $\text{Complex}(S)$ is a complexity measure.

If we think that fuzzy rules with “*not large*” are more complicated than other fuzzy rules with no negations, we may have the following order

$$\text{Complex}(S_1) > \text{Complex}(S_2) > \text{Complex}(S_3) > \text{Complex}(S_4). \quad (7)$$

B. Compatibility of Each Input Pattern with Its Winner Rule

Let us consider the classification of a new pattern $\mathbf{x}_1 = (0.05, 0.05)$ by the four fuzzy rule-based classifiers S_1 , S_2 , S_3 and S_4 . This pattern is classified as Class 2 by the following winner rule in each classifier:

R_1 in S_1 : If x_1 is *small* and x_2 is *small* then Class 2 with 1.0,

R_1 in S_2 : If x_1 is *small* and x_2 is *small* then Class 2 with 1.0,

R_{1245} in S_3 : If x_1 is *not large* and x_2 is *not large*
then Class 2 with 1.0,

R_5 in S_4 : If x_1 is *medium* and x_2 is *medium* then Class 2 with 1.0.

It seems that the fuzzy rule R_1 is a convincing explanation about the classification of $\mathbf{x}_1$ as Class 2. This is because $\mathbf{x}_1$ is linguistically described as “ x_1 is *small* and x_2 is *small*”, which means that $\mathbf{x}_1$ has high compatibility with the antecedent part of the fuzzy rule R_1 . The compatibility is calculated as

$$\mu_{A_1}(\mathbf{x}_1) = \mu_{\text{small}}(0.05) \times \mu_{\text{small}}(0.05) = 0.9 \times 0.9 = 0.81. \quad (8)$$

On the contrary, it seems that the fuzzy rule R_5 in S_4 is not a convincing explanation about the classification of $\mathbf{x}_1$ as Class 2. This is because $\mathbf{x}_1$ does not have high compatibility with the antecedent part of R_5 . Human users do not think that the input pattern $\mathbf{x}_1 = (0.05, 0.05)$ is well-described by the antecedent conditions “ x_1 is *medium* and x_2 is *medium*”. The compatibility of $\mathbf{x}_1$ with the antecedent part of R_5 is calculated as

$$\mu_{A_5}(\mathbf{x}_1) = \mu_{\text{medium}}(0.05) \times \mu_{\text{medium}}(0.05) = 0.1 \times 0.1 = 0.01. \quad (9)$$

From these discussions, we can see that the compatibility of the input pattern $\mathbf{x}_p$ with the antecedent part $A_{w(\mathbf{x}_p)}$ of its winner rule $R_{w(\mathbf{x}_p)}$ is related to the explanation ability of the fuzzy rule-based classifier. Let D_{Test} be the set of input patterns to be classified. This pattern set can be viewed as testing data that are not used for the design of the fuzzy rule-based classifier. The average compatibility of each input pattern with the antecedent part of its winner rule is calculated over the pattern set D_{Test} as follows:

$$\text{Compatibility}(S, D_{\text{Test}}) = \frac{\sum_{\mathbf{x}_p \in D_{\text{Test}}} \mu_{A_{w(\mathbf{x}_p)}}(\mathbf{x}_p)}{|D_{\text{Test}}|}, \quad (10)$$

where S is a fuzzy rule-based classifier, $|D_{\text{Test}}|$ is the cardinality of D_{Test} (i.e., the number of patterns in D_{Test}), and $A_{w(\mathbf{x}_p)}$ is the antecedent part of the winner rule $R_{w(\mathbf{x}_p)}$ for the input pattern $\mathbf{x}_p$ in D_{Test} . Since the winner rule $R_{w(\mathbf{x}_p)}$ is selected from the fuzzy rule-based classifier S , S is included in the left hand side whereas S is not explicitly shown in the right hand side.

The input pattern $\mathbf{x}_1 = (0.05, 0.05)$ is fully compatible with the antecedent part of the fuzzy rule R_{1245} in S_3 :

$$\mu_{A_{1245}}(\mathbf{x}_1) = \mu_{\text{not large}}(0.05) \times \mu_{\text{not large}}(0.05) = 1.0 \times 1.0 = 1.0. \quad (11)$$

Thus the fuzzy rule R_{1245} in S_3 has high explanation ability in terms of the compatibility with the input pattern. One may think that R_1 in S_1 seems to have higher explanation ability than R_{1245} in S_3 whereas R_{1245} has higher compatibility than R_1 . This intuition is related to the interpretability of each rule, which will be discussed later in the next section.

Let us consider another input pattern $\mathbf{x}_2 = (0.95, 0.05)$. This pattern is classified as Class 3 by the following winner rules:

R_7 in S_1 : If x_1 is *large* and x_2 is *small* then Class 3 with 1.0,

R_{789} in S_2 : If x_1 is *large* then Class 3 with 1.0,

R_{789} in S_3 : If x_1 is *large* then Class 3 with 1.0,

R_8 in S_4 : If x_1 is *large* and x_2 is *medium* then Class 3 with 1.0.

The fuzzy rule R_7 seems to be a convincing explanation about the classification of $\mathbf{x}_2$ because $\mathbf{x}_2$ is linguistically described as “ x_1 is *large* and x_2 is *small*”. The compatibility of $\mathbf{x}_2$ with the antecedent part of R_7 is calculated as

$$\mu_{A_7}(\mathbf{x}_2) = \mu_{\text{large}}(0.95) \times \mu_{\text{small}}(0.05) = 0.9 \times 0.9 = 0.81. \quad (12)$$

Whereas S_4 is simpler than S_1 , S_4 does not have higher explanation ability than S_1 . For example, the fuzzy rule R_8 in S_4 seems to be less convincing than R_7 in S_1 because “ x_1 is *large* and x_2 is *medium*” is not highly compatible with $\mathbf{x}_2 = (0.95, 0.05)$. The compatibility of $\mathbf{x}_2$ with the antecedent part of R_8 is calculated as

$$\mu_{A_8}(\mathbf{x}_2) = \mu_{\text{large}}(0.95) \times \mu_{\text{medium}}(0.05) = 0.9 \times 0.1 = 0.09. \quad (13)$$

On the contrary, the fuzzy rule R_{789} in S_2 and S_3 seems to be a good linguistic explanation. This is because its antecedent condition “ x_1 is *large*” describes $\mathbf{x}_2 = (0.95, 0.05)$ very well. The compatibility of $\mathbf{x}_2$ with the antecedent part of R_{789} is

$$\mu_{A_{789}}(\mathbf{x}_2) = \mu_{\text{large}}(0.95) = 0.90. \quad (14)$$

These discussions on the classification of $\mathbf{x}_2 = (0.95, 0.05)$ also suggest that the compatibility of the input pattern $\mathbf{x}_p$ with

the antecedent part $A_{w(x_p)}$ of its winner rule $R_{w(x_p)}$ is related to the explanation ability of the fuzzy rule-based classifier.

C. Interpretability of the Winner Rule for Each Input Pattern

In a fuzzy rule-based classifier with the single winner-based fuzzy reasoning method, the winner rule for an input pattern is used as the explanation for its classification result. Thus the interpretability of the winner rule is directly related to the explanation ability of the fuzzy rule-based classifier.

Let us consider the classification of the input pattern $x_1 = (0.05, 0.05)$ again. This pattern is classified by R_1 in S_1 :

R_1 in S_1 : If x_1 is *small* and x_2 is *small* then Class 2 with 1.0.

The fuzzy rule R_1 seems to have higher interpretability than the winner rule R_{1245} in S_3 :

R_{1245} in S_3 : If x_1 is *not large* and x_2 is *not large*
then Class 2 with 1.0.

This is because R_{1245} looks more complicated than R_1 .

When the same pattern x_1 is presented to the fuzzy rule-based classifier S_4 , it is classified as Class 2 by R_5 :

R_5 in S_4 : If x_1 is *medium* and x_2 is *medium* then Class 2 with 1.0.

This winner rule R_5 in S_4 can be viewed as having the same interpretability as R_1 in S_1 . This is because these two fuzzy rules have the same structure. As explained in the previous subsection, R_1 seems to have higher explanation ability for x_1 than R_5 because R_1 has higher compatibility. However, with respect to their interpretability, we cannot differentiate these two fuzzy rules.

Let us consider the other input pattern $x_2 = (0.95, 0.05)$. This pattern is classified as Class 3 by R_7 in S_1 and R_{789} in S_2 :

R_7 in S_1 : If x_1 is *large* and x_2 is *small* then Class 3 with 1.0,

R_{789} in S_2 : If x_1 is *large* then Class 3 with 1.0.

The interpretability of R_{789} seems to be higher than that of R_7 . This is because R_{789} has a simpler antecedent condition than R_7 . As in the case of the comparison between the two fuzzy rules R_1 and R_5 , the winner rule R_7 in S_1 can be viewed as having the same interpretability as the winner rule R_8 in S_4 :

R_8 in S_4 : If x_1 is *large* and x_2 is *medium* then Class 3 with 1.0.

Let $R_{w(x_p)}$ be the winner rule of the input pattern x_p . As in the previous subsection, let D_{Test} be the set of input patterns to be classified. In this case, the average interpretability of the winner rules over the pattern set D_{Test} is calculated as follows:

$$\text{Interpretability}(S, D_{\text{Test}}) = \frac{\sum_{x_p \in D_{\text{Test}}} \text{Interpretability}(R_{w(x_p)})}{|D_{\text{Test}}|} . \quad (15)$$

In (15), the problem is how to define the interpretability of each fuzzy rule. In the above discussions, we intuitively use the

complexity of the antecedent part of each fuzzy rule. A simple measure to evaluate the complexity of the antecedent part is the number of antecedent conditions. This measure seems to be too simple because it cannot differentiate between R_{1245} with the antecedent condition “If x_1 is *not large* and x_2 is *not large*” and R_1 with the antecedent condition “If x_1 is *small* and x_2 is *small*”.

We may be able to refer to a lot of existing studies on the interpretability of fuzzy rule-based systems [11]–[19]. It should be noted, however, that the interpretability in those studies was usually discussed for fuzzy rule-based classifiers rather than each individual fuzzy rule. In (15), the interpretability of each winner rule is used. The interpretability in (15) can be viewed as the understandability of the winner rule for human users.

D. Accuracy of the Winner Rule for Each Input Pattern

In multiobjective fuzzy rule-based classifier design, the accuracy of classifiers is not viewed as a factor related to the interpretability. This is because the accuracy is handled as a separate objective from the interpretability in multiobjective fuzzy rule-based classifier design. However, the accuracy of the winner rule seems to be an important factor related to the explanation ability of fuzzy rule-based classifiers.

Let us consider the following fuzzy rule-based classifier with three rules for our three-class test problem:

S_5 : Fuzzy rule-based classifier with three fuzzy rules

R_{123} : If x_1 is *small* then Class 1 with 1.0,

R_{456} : If x_1 is *medium* then Class 2 with 1.0,

R_{789} : If x_1 is *large* then Class 3 with 1.0.

The classification boundaries by S_5 are shown in Fig. 4. Several training patterns from Class 1 and Class 2 are misclassified while all the 20 patterns from Class 3 are correctly classified in Fig. 4.

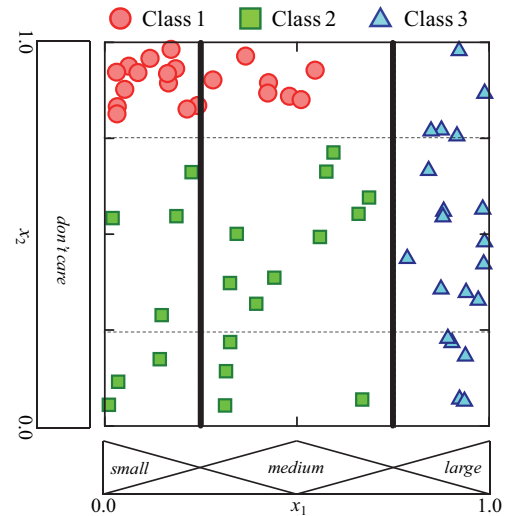


Fig. 4. Classification boundaries by S_5 .

As in the previous two subsections, let us consider the classification of the input pattern $x_1 = (0.05, 0.05)$. This pattern is classified as Class 1 by the winner rule R_{123} in S_5 :

R_{123} in S_5 : If x_1 is *small* then Class 1 with 1.0.

This winner rule R_{123} has high compatibility with $\mathbf{x}_1 = (0.05, 0.05)$. This winner rule has also has high interpretability thanks to its simple antecedent condition. As a result, this winner rule is viewed as having high explanation ability with respect to the compatibility and the interpretability. However, we do not think that the winner rule R_{123} explains the classification result of $\mathbf{x}_1$ in a convincing manner. The winner rule R_{123} classifies $\mathbf{x}_1 = (0.05, 0.05)$ as Class 1. However, this classification result is not likely to be correct in Fig. 4 because many training patterns close to $\mathbf{x}_1$ in Fig. 4 are from Class 2. That is, we may think that the explanation ability of R_{123} is low since the accuracy of R_{123} is low.

From these discussions, we can see that the accuracy of the winner rule is related to the explanation ability of a fuzzy rule-based classifier. The average accuracy of the winner rule $R_{w(\mathbf{x}_p)}$ can be defined over the pattern set D_{Test} as follows:

$$\text{Accuracy}(S, D_{\text{Test}}) = \frac{\sum_{\mathbf{x}_p \in D_{\text{Test}}} \text{Accuracy}(R_{w(\mathbf{x}_p)})}{|D_{\text{Test}}|}. \quad (16)$$

In (16), the problem is the definition of the accuracy of each fuzzy rule. One candidate definition is as follows:

$$\text{Accuracy}(R_q) = \frac{\text{Number of patterns correctly classified by } R_q}{\text{Number of patterns classified by } R_q}. \quad (17)$$

Let us assume that a single fuzzy rule correctly classifies 18 out of 20 patterns for which it is selected as the winner rule. The accuracy of this rule is calculated as 18/20 (i.e., 0.9) in (17). This accuracy is summed up 20 times in the numerator of (16), which is calculated as 18 (i.e., the same as the number of correctly classified patterns by this fuzzy rule). From this discussion, we can see that (16) is the same as the classification rate when we use (17) as the accuracy of each fuzzy rule:

$$\begin{aligned} \text{Accuracy}(S, D_{\text{Test}}) \\ = \frac{\text{Number of patterns correctly classified by } S}{|D_{\text{Test}}|}. \end{aligned} \quad (18)$$

In this definition based on (16) and (17), the accuracy of the fuzzy rule-based classifier S is directly used as a factor related to its explanation ability.

The accuracy of the fuzzy rule R_q can be also defined as

$$\text{Accuracy}(R_q) = \frac{\sum_{\mathbf{x}_p \in D_{\text{Test}}(C_q)} \mu_{A_q}(\mathbf{x}_p)}{\sum_{\mathbf{x}_p \in D_{\text{Test}}} \mu_{A_q}(\mathbf{x}_p)}, \quad (19)$$

where $D_{\text{Test}}(C_q)$ is the set of patterns in D_{Test} from Class C_q (i.e., the same class as the consequent class of the fuzzy rule R_q). In the field of fuzzy data mining [29]-[32], the accuracy in (19) is referred to as the confidence of the fuzzy rule R_q .

In (19), it is possible to use the training data rather than the

test data. In data mining, the confidence is usually defined using the training data. If we use the 60 patterns in Fig. 4 as the data set D_{Test} in (19), the accuracy (i.e., confidence) of each rule in S_5 is calculated from (19) as follows:

$$\begin{aligned} \text{Accuracy}(R_{123}) &= \text{Confidence}(R_{123}) = 0.591, \\ \text{Accuracy}(R_{456}) &= \text{Confidence}(R_{456}) = 0.469, \\ \text{Accuracy}(R_{789}) &= \text{Confidence}(R_{789}) = 0.911. \end{aligned}$$

If we also use the 60 patterns in Fig. 4 as the data set D_{Test} in (16), the accuracy measure of the fuzzy rule-based classifier S_5 is calculated as follows:

$$\text{Accuracy}(S_5, D_{\text{Test}}) = 0.657.$$

For comparison, we calculate the accuracy measure for each of the other fuzzy rule-based classifiers S_1, S_2, S_3 and S_4 in the same manner as follows:

$$\begin{aligned} \text{Accuracy}(S_1, D_{\text{Test}}) &= 0.846, \\ \text{Accuracy}(S_2, D_{\text{Test}}) &= 0.848, \\ \text{Accuracy}(S_3, D_{\text{Test}}) &= 0.837, \\ \text{Accuracy}(S_4, D_{\text{Test}}) &= 0.784. \end{aligned}$$

We can see that S_1, S_2, S_3 and S_4 with no misclassified pattern have larger values of the accuracy measure than S_5 with many misclassified patterns.

E. Coverage of the Winner Rule for Each Input Pattern

Roughly speaking, fuzzy rules with small covering regions in the pattern space tend to have high accuracy. Let us consider an extreme case where each fuzzy rule covers only a single pattern. In this extreme case, the accuracy of each fuzzy rule is always 1.0. However, such a fuzzy rule with a very small covering region may have small explanation ability.

Let us consider the following two cases:

Case 1: Most patterns are classified by a few winner rules. That is, many patterns share the same winner rule.

Case 2: Each pattern is classified by a different winner rule. That is, each winner rule classifies only a single pattern.

We think that Case 1 has higher explanation ability than Case 2. This is because human users do not have to examine many winner rules in Case 1.

In fuzzy data mining, coverage is defined as follows:

$$\text{Coverage}(R_q) = \frac{\sum_{\mathbf{x}_p \in D_{\text{Test}}(C_q)} \mu_{A_q}(\mathbf{x}_p)}{|D_{\text{Test}}(C_q)|}, \quad (20)$$

where $|D_{\text{Test}}(C_q)|$ is the number of D_{Test} from Class C_q .

We can also define the coverage of R_q using the number of correctly classified patterns as

$$\text{Coverage}(R_q) = \frac{\text{Number of correctly classified patterns by } R_q}{|D_{\text{Test}}(C_q)|}. \quad (21)$$

We can see that the coverage of a fuzzy rule in (21) is large when it correctly classifies many patterns (i.e., when it is selected as the winner rule for many patterns from its consequent class). The coverage in (20) has a similar meaning: The coverage of a fuzzy rule in (20) has a large value when it covers many patterns from its consequent class.

The average value of coverage of the winner rule is calculated over the pattern set D_{Test} as follows:

$$\text{Coverage}(S, D_{\text{Test}}) = \frac{\sum_{x_p \in D_{\text{Test}}} \text{Coverage}(R_{w(x_p)})}{|D_{\text{Test}}|} . \quad (22)$$

If we use the 60 patterns in Fig. 4 as the pattern set D_{Test} , the average value of coverage in (22) is calculated for each fuzzy rule-based classifier for the definition in (21) as follows:

$$\begin{aligned} \text{Coverage}(S_1, D_{\text{Test}}) &= 0.408, \\ \text{Coverage}(S_2, D_{\text{Test}}) &= 0.617, \\ \text{Coverage}(S_3, D_{\text{Test}}) &= 1.000, \\ \text{Coverage}(S_4, D_{\text{Test}}) &= 0.875, \\ \text{Coverage}(S_5, D_{\text{Test}}) &= 0.767. \end{aligned}$$

We can see that high coverage is obtained from fuzzy rule-based classifiers with a small number of fuzzy rules (e.g., S_3 and S_5 with three fuzzy rules, and S_4 with four fuzzy rules). Since many patterns are misclassified by S_5 , its coverage by (21) is much smaller than S_3 whereas they have the same number of fuzzy rules.

If we use the definition of (20), the average value of coverage in (22) is calculated for each fuzzy rule-based classifier as follows:

$$\begin{aligned} \text{Coverage}(S_1, D_{\text{Test}}) &= 0.326, \\ \text{Coverage}(S_2, D_{\text{Test}}) &= 0.494, \\ \text{Coverage}(S_3, D_{\text{Test}}) &= 0.875, \\ \text{Coverage}(S_4, D_{\text{Test}}) &= 0.364, \\ \text{Coverage}(S_5, D_{\text{Test}}) &= 0.641. \end{aligned}$$

We can also see that high coverage is obtained from fuzzy rule-based classifiers with a small number of fuzzy rules (e.g., S_3 and S_5 with three fuzzy rules). Since many patterns are misclassified by S_5 , its coverage by (20) is much smaller than S_3 whereas they have the same number of fuzzy rules.

IV. CONCLUSIONS

In this paper, first we defined the explanation ability as “the ability of a fuzzy rule-based classifier to explain why an input pattern is classified as a particular class in a convincing way.” Then we pointed out the following four factors related to the quantitative definition (i.e., mathematical formulation) of the explanation ability:

- (1) Compatibility of an input pattern with its winner rule
- (2) Interpretability of the winner rule
- (3) Accuracy of the winner rule
- (4) Coverage of the winner rule

We discussed the formulation of each factor. It was clearly shown through the discussions on these four factors that the explanation ability of a fuzzy rule-based classifier is related to both its accuracy and its interpretability. Since we formulated each factor, we may be able to formulate the explanation ability by combining those formulations into a single overall measure. A discussion on how to combine them into a single overall measure is a future research issue. This issue does not exclude the possibility of combining the four factors into two measures with different characteristic features.

We may need further discussions on each factor as well as the above-mentioned discussions on their integration. It may be interesting to examine relations among the four factors. We also need further discussions on the possibility of other factors that are related to the explanation ability. Moreover, it may be a promising research issue to use the explanation ability instead of or together with the interpretability in multiobjective fuzzy rule-based classifier design. In this case, both of the classifier-level accuracy and the rule-level accuracy will be used to measure different aspects of fuzzy rule-based classifiers.

REFERENCES

- [1] H. Ishibuchi, K. Nozaki, N. Yamamoto, and H. Tanaka, “Selecting fuzzy if-then rules for classification problems using genetic algorithms,” *IEEE Trans. on Fuzzy Systems* 3 (3) 260-270, 1995.
- [2] M. Setnes and H. Roubos, “GA-fuzzy modeling and classification: Complexity and performance,” *IEEE Trans. on Fuzzy Systems* 8 (5) 509-522, 2000.
- [3] L. Castillo, A. González, and R. Pérez, “Including a simplicity criterion in the selection of the best rule in a genetic fuzzy learning algorithm,” *Fuzzy Sets and Systems* 120 (2) 309-321, 2001.
- [4] J. Abonyi, J. A. Roubos, and F. Szeifert, “Data-driven generation of compact, accurate, and linguistically sound fuzzy classifiers based on a decision-tree initialization,” *International Journal of Approximate Reasoning* 32 (1) 1-21, 2003.
- [5] 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 and Systems* 89 (2) 135-150, 1997.
- [6] H. Ishibuchi, T. Nakashima, and T. Murata, “Three-objective genetics-based machine learning for linguistic rule extraction,” *Information Sciences* 136 (1-4) 109-133, 2001.
- [7] P. Pulkkinen and H. Koivisto, “Fuzzy classifier identification using decision tree and multiobjective evolutionary algorithms,” *International Journal of Approximate Reasoning* 48 (2) 526-543, 2008.
- [8] P. Ducange, B. Lazzerini, and F. Marcelloni, “Multi-objective genetic fuzzy classifiers for imbalanced and cost-sensitive datasets,” *Soft Computing* 14 (7) 713-728, May 2010.
- [9] Y. Zhang, X. B. Wu, Z. Y. Xing, and W. L. Hu, “On generating interpretable and precise fuzzy systems based on Pareto multi-objective cooperative co-evolutionary algorithm,” *Applied Soft Computing* 11 (1) 1284-1294, 2011.
- [10] H. Ishibuchi and Y. Nojima, “Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning,” *International Journal of Approximate Reasoning* 44 (1) 4-31, 2007.
- [11] J. Casillas, O. Cordón, F. Herrera, and L. Magdalena (eds.), *Interpretability Issues in Fuzzy Modeling*, Springer, 2003.
- [12] J. Casillas, O. Cordón, F. Herrera, and L. Magdalena (eds.), *Accuracy Improvements in Linguistic Fuzzy Modeling*, Springer, 2003.
- [13] S. Guillaume, “Designing fuzzy inference systems from data: An interpretability-oriented review,” *IEEE Trans. on Fuzzy Systems* 9 (3) 426-443, 2001.

- [14] R. Mikut, J. Jäkel, and L. Gröll, "Interpretability issues in data-based learning of fuzzy systems," *Fuzzy Sets and Systems* 150 (2) 179-197, 2005.
- [15] C. Mencar and A. M. Fanelli, "Interpretability constraints for fuzzy information granulation," *Information Sciences* 178 (24) 4585-4618, 2008.
- [16] 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 and Systems* 159 (23) 3091-3131, 2008.
- [17] A. Botta, B. Lazzerini, F. Marcelloni, and D. C. Stefanescu, "Context adaptation of fuzzy systems through a multi-objective evolutionary approach based on a novel interpretability index," *Soft Computing* 13 (5) 437-449, 2009.
- [18] J. M. Alonso, L. Magdalena, and G. González-Rodríguez, "Looking for a good fuzzy system interpretability index: An experimental approach," *International Journal of Approximate Reasoning*, 51 (1) 115-134, 2009.
- [19] C. Mencar, C. Castiello, R. Cannone, and A. M. Fanelli, "Interpretability assessment of fuzzy knowledge bases: A cointension based approach," *International Journal of Approximate Reasoning* 52 (4) 501-518, June 2011.
- [20] H. Ishibuchi, Y. Kaisho, and Y. Nojima, "Designing fuzzy rule-based classifiers that can visually explain their classification results to human users," *Proc. of 3rd International Workshop on Genetic and Evolving Fuzzy Systems*, 5-10, 2008.
- [21] H. Ishibuchi, Y. Kaisho, and Y. Nojima, "A visual explanation system for explaining fuzzy reasoning results by fuzzy rule-based classifiers," *Proc. of 2008 North American Fuzzy Information Processing Society Conference*, CD-ROM Proceeding (6 pages), 2008.
- [22] H. Ishibuchi and Y. Nojima, "Discussions on Interpretability of Fuzzy Systems using Simple Examples," *Proc. of 13th IFSA World Congress and 6th Conference of EUSFLAT*, 1649-1654, 2009.
- [23] H. Ishibuchi, Y. Kaisho, and Y. Nojima, "Complexity, interpretability and explanation capability of fuzzy rule-based classifiers," *Proc. of 2009 IEEE International Conference on Fuzzy Systems*, 1730-1735, Jeju Island, Korea, August 20-24, 2009.
- [24] H. Ishibuchi, Y. Kaisho, and Y. Nojima, "Design of linguistically interpretable fuzzy rule-based classifiers: A short review and open questions," *Journal of Multiple-Valued Logic and Soft Computing* 17 (2-3) 101-134, 2011.
- [25] H. Ishibuchi and T. Yamamoto, "Rule weight specification in fuzzy rule-based classification systems," *IEEE Trans. on Fuzzy Systems* 13 (4) 428-435, 2005.
- [26] K. Nozaki, H. Ishibuchi, and H. Tanaka, "Adaptive fuzzy rule-based classification systems," *IEEE Trans. on Fuzzy Systems* 4 (3) 238-250, 1996.
- [27] O. Cordón, M. J. del Jesus, and F. Herrera, "A proposal on reasoning methods in fuzzy rule-based classification systems," *International Journal of Approximate Reasoning* 20 (1) 21-45, 1999.
- [28] H. Ishibuchi, T. Nakashima, and T. Morisawa, "Voting in fuzzy rule-based systems for pattern classification problems," *Fuzzy Sets and Systems* 103 (2) 223-238, 1999.
- [29] H. Ishibuchi, T. Nakashima, and M. Nii: *Classification and Modeling with Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining*, Springer, Berlin, 2004.
- [30] H. Ishibuchi, T. Yamamoto, and T. Nakashima, "Fuzzy data mining: Effect of fuzzy discretization," *Proc. of 2001 IEEE International Conference on Data Mining*, 241-248, 2001.
- [31] T. P. Hong, C. S. Kuo, and S. C. Chi, "Trade-off between computation time and number of rules for fuzzy mining from quantitative data," *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems* 9 (5) 587-604, 2001.
- [32] T. P. Hong, K. Y. Lin, and S. L. Wang, "Fuzzy data mining for interesting generalized association rules," *Fuzzy Sets and Systems* 138 (2) 255-269, 2003.