The Thirteenth International Symposium on Artificial Life and Robotics 2008(AROB 13th '08),
B-Con Plaza, Beppu, Oita, Japan, January 31-February 2, 2008

# Effects of Constructing Fuzzy Discretization from Crisp Discretization for Rule-based Classifiers

Isao Kuwajima, Yusuke Nojima, and Hisao Ishibuchi

*Gakuen-cho 1-1, Naka-ku, Sakai, Osaka 599-8531, Japan*
*(Tel : 81-072-254-9198; Fax : 81-072-254-9915)*
*(Email {kuwajima@ci., nojima@, hisaoi@}cs.osakafu-u.ac.jp)*

***Abstract***: Crisp discretization is one of the most widely used methods for handling continuous attributes. In crisp discretization, each attribute is split into several intervals and handled as discrete numbers. Although crisp discretization is a convenient tool, it is not appropriate in some situations (e.g., when there is no clear boundary and we cannot set a clear threshold). To address such a problem, several discretizations with fuzzy sets have been proposed. In this paper we examine the effect of constructing fuzzy discretization from crisp discretization. The fuzziness of fuzzy discretization is controlled by a fuzzification grade $F$. We examine two procedures for the setting of $F$. In one procedure, we set $F$ beforehand and do not change it through training rule-based classifiers. In the other procedure, first we set $F$ and then change it after training. Through computational experiments, we show that the accuracy of rule-based classifiers is improved by an appropriate setting of the grade of fuzzification. Moreover, we show that increasing the grade of fuzzification after training classifiers can often improve generalization ability.

***Keywords***: Fuzzy sets, genetic rule selection, pattern classification.

## I. INTRODUCTION

When we classify patterns with continuous attributes by rule-based classifiers, each attribute is usually split into intervals. This process of transformation is called crisp discretization. Crisp discretization has been used in numerous machine learning techniques such as decision trees. In some situations, crisp discretization is appropriate for handling continuous attributes. In other situations, however, it is not appropriate. For example, we usually use linguistic terms for dividing our height into some categories (e.g., short, medium, and tall). Such a linguistic term cannot be represented appropriately by crisp discretization. Using fuzzy sets, we can represent linguistic terms. Fuzzy discretization can avoid certain undesirable threshold effects. It has been shown that while crisp discretization can be fine-tuned to training patterns, it often leads to the overfitting and low generalization ability (i.e., low accuracy on test patterns) [1]. It has also been shown that fuzzy discretization often has a smoother decision boundary and higher generalization ability [1].

In this paper, we examine the effects of constructing fuzzy discretization from crisp discretization. The fuzziness of fuzzy discretization is controlled by a fuzzification grade $F$. We examine two procedures for the setting of $F$. In one procedure, we set $F$ beforehand and do not change it through training classifiers. In the other procedure, first we set $F$ and change it after training. Through computational experiments, we show that the generalization ability of classifiers is improved by appropriate settings of $F$. Moreover, we show that increasing the fuzzification grade after training classifiers can often improve generalization ability.

## II. FUZZY DISCRETIZATION FROM CRISP DISCRETIZATION

The idea and method of fuzzy discretization from crisp discretization was first proposed in [2]. It is constructed as follows:

(a) Membership functions are linear (i.e., triangular or trapezoidal).
(b) The sum of the membership values of neighboring fuzzy sets is 1.
(c) Crossing points of neighboring membership functions are identical to threshold values for intervals.
(d) The membership values of intermediate fuzzy set in the domain interval [0, 1] (e.g., MS: *medium small*, M: *medium*, and ML: *medium large* in Fig. 1) is 1 at the midpoint of the corresponding intervals. The membership value of the smallest fuzzy set (e.g., S: *small* in Fig. 1) is 1 at the smallest input value 0 in the domain interval [0, 1]. The membership value of the largest fuzzy set (e.g., L: *large* in Fig. 1.) is 1 at the largest input value 1 in the domain interval [0, 1].

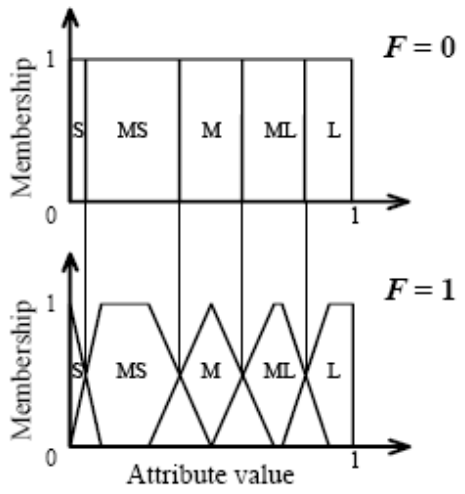



Fig. 1. Crisp discretization with $F = 0$ (top) and fuzzy discretization with $F = 1$ (bottom).

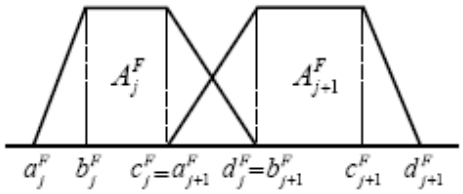



Fig. 2. Adjacent trapezoidal fuzzy sets.

Figure 1 is an example of fuzzy discretization constructed from crisp discretization. It should be noted that fuzzy discretization cannot be uniquely specified by the above four constraint conditions. To specify fuzzy discretization uniquely, let us consider the fuzzification grade $F$ of fuzzy discretization. When $F = 1$, fuzzy discretization is fully fuzzified under the above constraint conditions as shown in Fig. 1. In contrast, $F = 0$ corresponds to crisp discretization which has no overlaps between adjacent fuzzy sets. From fuzzy discretization with $F = 1$ and crisp discretization with $F = 0$, we can generate partially fuzzified discretization with arbitrary grades of fuzzification.

Let us denote partially fuzzified trapezoidal fuzzy set with fuzzification grade $F$ as $A_j^F = (a_j^F, b_j^F, c_j^F, d_j^F)$ where $0 \le F \le 1$ (see Fig. 2). Note that $A_j^0$ and $A_j^1$ correspond to crisp discretization and fully fuzzified discretization, respectively. Using the interpolation between $A_j^0$ and $A_j^1$, we can specify $A_j^F = (a_j^F, b_j^F, c_j^F, d_j^F)$ as follows:

$$a_j^F = a_j^0 + (a_j^1 + a_j^0) \cdot F, \tag{1}$$

$$b_j^F = b_j^0 + (b_j^1 + b_j^0) \cdot F, \tag{2}$$

$$c_j^F = c_j^0 + (c_j^1 + c_j^0) \cdot F, \tag{3}$$

$$d_j^F = d_j^0 + (d_j^1 + d_j^0) \cdot F. \tag{4}$$

## III. RULE-BASED CLASSIFIERS

Let $\mathbf{x} = (x_1, ..., x_n)$ be an $n$-dimensional pattern vector. Fuzzy rules for an $n$-dimensional classification problem are written as,

$$\begin{aligned} \text{Rule } R_q : \text{If } x_1 \text{ is } A_{q1} \text{ and } ... \text{ and } x_n \text{ is } A_{qn} \\ \text{then Class } C_q \text{ with } CF_q, \end{aligned} \tag{5}$$

where $\mathbf{A}_q = (A_{q1}, ..., A_{qn})$ is an antecedent set, $C_q$ is a consequent class, and $CF_q$ is a rule weight (i.e., certainty factor). We denote the rule $R_q$ in (5) as $\mathbf{A}_q \Rightarrow C_q$. It should be noted that $R_q$ in (5) does not always have $n$ antecedent conditions. Some conditions can be *don't care*.

The membership value of the pattern $\mathbf{x}$ to the antecedent set $\mathbf{A}_q$ is calculated with the product operator as

$$\mu_{\mathbf{A}_q}(\mathbf{x}) = \mu_{A_{q1}}(x_1) \cdot ... \cdot \mu_{A_{qn}}(x_n), \tag{6}$$

where $\mu_{A_{qi}}(\cdot)$ is the membership function of the antecedent fuzzy set $A_{qi}$.

In the field of data mining, two rule evaluation criteria called *confidence* and *support* are often used to examine the quality of a rule. Let us assume that we have $m$ training patterns $\mathbf{x}_p$, $p = 1, 2, ..., m$. Then the fuzzy version of the confidence is defined as follows [3]:

$$c(\mathbf{A}_q \Rightarrow \text{Class } h) = \frac{\sum_{\mathbf{x}_p \in \text{Class } h} \mu_{\mathbf{A}_q}(\mathbf{x}_p)}{\sum_{p=1}^{m} \mu_{\mathbf{A}_q}(\mathbf{x}_p)}. \tag{7}$$

In the same manner, the support is defined as follows:

$$s(\mathbf{A}_q \Rightarrow \text{Class } h) = \frac{\sum_{\mathbf{x}_p \in \text{Class } h} \mu_{\mathbf{A}_q}(\mathbf{x}_p)}{m}. \tag{8}$$

By using the confidence measure, the rule weight $CF_q$ is specified as [4]

$$CF_q = c(\mathbf{A}_q \Rightarrow \text{Class } C_q) - \sum_{\substack{h=1 \\ h \neq C_q}}^{M} c(\mathbf{A}_q \Rightarrow \text{Class } h). \tag{9}$$

Let $S$ be a set of rules of the form (5). That is, $S$ is a rule-based classifier. A new pattern $\mathbf{x}$ is classified by a single winner rule $R_w$ which is chosen from the rule set $S$ as follows:

$$\mu_{\mathbf{A}_w}(\mathbf{x}_p) \cdot CF_w = \max\{\mu_{\mathbf{A}_q}(\mathbf{x}_p) \cdot CF_q \mid R_q \in S\}. \quad (10)$$

## IV. EVOLUTIONARY MULTIOBJECTIVE RULE SELECTION

In order to obtain simple and accurate classifiers, we use evolutionary multiobjective rule selection as in [4]. Let us assume that we have already extracted $N$ rules by data mining techniques. A subset $S$ of the $N$ candidate rules is represented by a binary string of length $N$ as

$$S = s_1 s_2 \ldots s_N, \quad (11)$$

where $s_j = 1$ and $s_j = 0$ mean that the $j$-th candidate rule is included in $S$ and excluded from $S$, respectively.

In order to obtain simple and accurate classifiers, the following three objectives are optimized:

$f_1(S)$: The number of correctly classified patterns,
$f_2(S)$: The number of rules,
$f_3(S)$: The total number of antecedent fuzzy sets.

The first objective is maximized while the second and third objectives are minimized.

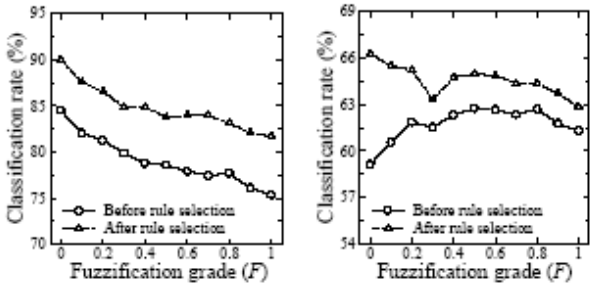
## V. COMPUTATIONAL EXPERIMENTS

Our rule-based classifiers with fuzzy discretization were evaluated on the following eight data sets from the UCI machine learning repository: Breast W, Diabetes, Glass, Heart C, Iris, Liver, Sonar, and Wine. For all of these data sets, the accuracy of classifiers was examined by iterating the two-fold cross-validation five times ($5 \times 2$CV). We specified the minimum confidence and support as 0.7 and 0.04, and chose $\varepsilon$-Pareto-optimal rules [4] with $\varepsilon = 0.04$ as candidate rules. Crisp discretization was constructed by dividing the domain of each attribute by the optimal splitting method [5] based on the class entropy measure. We examine two procedures for the setting of $F$ in the following subsections.

### 1. Setting the Grade of Fuzzification before Rule Selection

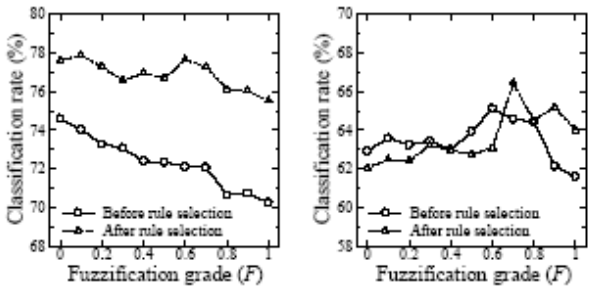
First we examine the effects of fuzzification grade $F$ on the accuracy. We set $F$ beforehand and do not change it through training rule-based classifiers. Figures 3 and 4 show the relation between the setting of $F$ and average accuracy. Average accuracy was calculated with classifiers that have the highest accuracy on training data. From Figs. 3 and 4, it can be seen that the larger $F$ values lead to less training accuracy. However, the larger $F$ values do not always lead to less test accuracy.

To show this more clearly, we examine the correlation between $F$ and accuracy. Table 1 shows the correlation coefficient between $F$ and accuracy. From Table 1, we can see that there are strong negative correlations between $F$ and training accuracy. In contrast, there is no consistent correlation between $F$ and test accuracy.




(a) Training data. (b) Test data.

Fig. 3. Relation between accuracy and $F$ values (Glass).




(a) Training data. (b) Test data.

Fig. 4. Relation between accuracy and $F$ values (Liver).

Table 1. Correlation coefficients between $F$ values and accuracy on training and test data.

| | Training data | | Test data | |
|---|---|---|---|---|
| Data set | Before | After | Before | After |
| Breast W | -0.94 | -0.91 | 0.95 | 0.04 |
| Diabetes | -0.98 | -0.91 | -0.63 | 0.11 |
| Glass | -0.96 | -0.94 | 0.56 | -0.75 |
| Heart C | -0.97 | -0.97 | -0.49 | 0.63 |
| Iris | -0.96 | 0.56 | 0.60 | -0.38 |
| Liver | -0.98 | -0.76 | -0.13 | 0.75 |
| Sonar | -0.75 | -0.50 | 0.97 | 0.82 |
| Wine | -0.85 | N/A | 0.99 | 0.86 |

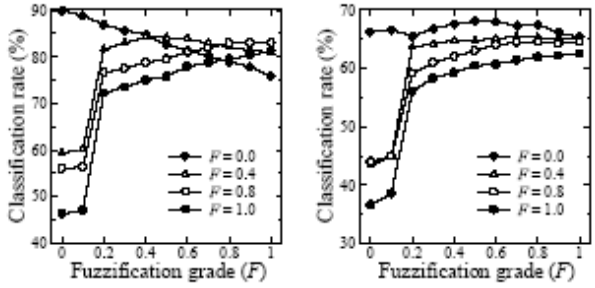
Before: before rule selection, After: after rule selection.

### 2. Changing the Grade of Fuzzification after Rule Selection

While we specified the fuzzification grade $F$ before rule selection in the previous subsection, it can be

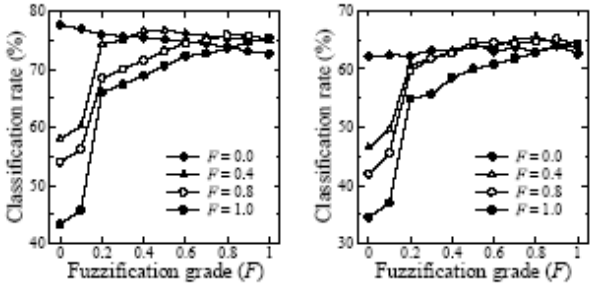
changed at any time. In this subsection, we examine the effects of changing the fuzzification grade $F$ after rule selection. While the fuzzification grade $F$ was changed after rule selection in the computational experiments, the rule weight $CF$ and other parameters was not changed.

Figures 5 and 6 show the results of changing the grade of fuzzification after rule selection for four initial values of $F$ ($F$ = 0.0, 0.4, 0.8, 1.0). It can be seen that when the initial value of $F$ is large (e.g., $F$ = 1), decreasing the fuzzification grade often leads to low accuracy. It can be also seen that when the initial value of $F$ is small (e.g., $F$ = 0, 0.4), increasing the fuzzification grade leads to slightly better test accuracy.




(a) Training data. (b) Test data.

Fig. 5. Accuracy when the $F$ value is changed after training (Glass).




(a) Training data. (b) Test data.

Fig. 6. Accuracy when the $F$ value is changed after training (Liver).

We examine the case where the initial value of $F$ is 0 (i.e., crisp discretization) because it is widely used in machine learning. Table 2 shows the results of changing the grade of fuzzification after training for the crisp discretization (i.e., $F$ = 0). For all data set, increasing the $F$ value from zero led to the better test accuracy. From these results, if rule-based classifiers are obtained with crisp discretization, it is a reasonable choice to increase the fuzzification grade $F$ after training.

Table 2. Test accuracy (%) when the $F$ is changed after training for crisp discretization (initial value of $F$ is 0).

| | Fuzzification grade $F$ | | | | | |
|---|---|---|---|---|---|---|
| Data set | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1 |
| Breast W | 94.6 | 94.6 | 94.8 | 94.8 | 95.2 | **95.4** |
| Diabetes | 73.7 | **74.0** | 73.3 | 72.8 | 72.3 | 72.3 |
| Glass | 66.3 | 65.6 | 67.6 | **68.1** | 67.5 | 65.3 |
| Heart C | 52.4 | 53.0 | 53.6 | 53.8 | 54.5 | **54.9** |
| Iris | 94.4 | 94.8 | 95.2 | 95.2 | 95.4 | **95.8** |
| Liver | 62.1 | 62.3 | **63.0** | 62.9 | **63.0** | 62.6 |
| Sonar | 67.9 | 68.1 | 68.9 | 69.2 | 70.1 | **71.3** |
| Wine | 89.3 | 90.3 | 92.1 | 93.3 | **93.5** | 92.1 |

## VI. CONCLUSION

In this paper, we examined the effects of constructing fuzzy discretization from crisp discretization. We examined two procedures for the setting of the fuzzification grade. Through experiments, we showed that the generalization ability of classifiers was improved by an appropriate setting of fuzzification grade. Moreover, we showed that increasing the fuzzification grade after rule selection was a promising approach for improving classifiers' generalization ability.

## REFERENCES

[1] Ishibuchi H, Nojima Y (2005), Comparison between fuzzy and interval partitions in evolutionary multiobjective design of rule-based classification systems, Proc. of 2005 IEEE International Conference on Fuzzy Systems: 430-435.

[2] Ishibuchi H, Yamamoto T (2002), Performance evaluation of fuzzy partitions with different fuzzification grades, Proc. of IEEE International Conference on Fuzzy Systems: 1198-1203

[3] Hong TP, Kuo CS, Chi SC (2001), Tradeoff between computation time and number of rules for fuzzy mining from quantitative data, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9 (6): 587-604.

[4] Ishibuchi H, Kuwajima I, Nojima Y (2007), Relation between Pareto-optimal fuzzy rules and Pareto-optimal fuzzy rule sets, Proc. of IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making: 42-49.

[5] Elomaa T, Rousu J (1999), General and efficient multisplitting of numerical attributes, Machine Learning 36 (3): 201-244.