Relation between Pareto-Optimal Fuzzy Rules and Pareto-Optimal Fuzzy Rule Sets

Hisao Ishibuchi, Isao Kuwajima, and Yusuke Nojima
Department of Computer Science and Intelligent Systems, Osaka Prefecture University
hisaoi@cs.osakafu-u.ac.jp, kuwajima@ci.cs.osakafu-u.ac.jp, nojima@cs.osakafu-u.ac.jp

Abstract – Evolutionary multiobjective optimization (EMO) has been utilized in the field of data mining in the following two ways: to find Pareto-optimal rules and Pareto-optimal rule sets. Confidence and coverage are often used as two objectives to evaluate each rule in the search for Pareto-optimal rules. Whereas all association rules satisfying the minimum support and confidence are usually extracted in data mining, only Pareto-optimal rules are searched for by an EMO algorithm in multiobjective data mining. On the other hand, accuracy and complexity are used to evaluate each rule set. The complexity of each rule set is often measured by the number of rules and the number of antecedent conditions. An EMO algorithm is used to search for Pareto-optimal rule sets with respect to accuracy and complexity. In this paper, we examine the relation between Pareto-optimal rules and Pareto-optimal rule sets in the design of fuzzy rule-based systems for pattern classification problems. More specifically, we check whether Pareto-optimal rules are included in Pareto-optimal rule sets through computational experiments using multiobjective genetic fuzzy rule selection. A mixture of Pareto-optimal and non Pareto-optimal fuzzy rules are used as candidate rules in multiobjective genetic fuzzy rule selection. We also examine the performance of selected rules when we use only Pareto-optimal rules as candidate rules.

I. INTRODUCTION

Recently association rule mining techniques have been applied to classification problems to design an accurate and compact classifier [1]-[4]. In these studies, the design of classifiers is performed through the following two steps: rule discovery and rule pruning. A large number of candidate classification rules are extracted from numerical data in the rule discovery step. An association rule mining technique such as Apriori [5] is used to find all rules satisfying the minimum support and confidence. Whereas these two rule evaluation measures (i.e., support and confidence) have been frequently used in data mining, other measures were also proposed to evaluate association rules. Among them are gain, variance, chi-squared value, entropy gain, gini, laplace, lift, and conviction [6]. It was shown for non-fuzzy rules that the best rule according to any of these measures is a Pareto-optimal rule of a two-objective rule discovery problem with support maximization and confidence maximization [6]. The use of an evolutionary multiobjective optimization (EMO) algorithm was proposed to search for Pareto-optimal rules of this two-objective problem for partial classification [7]-[10]. Partial classification is the classification of a particular class (usually minor class) from all the other classes. Similar multiobjective formulations to [7]-[10] were used to search for Pareto-optimal association rules [11] and Pareto-optimal fuzzy association rules [12].

In the rule pruning step, a small number of rules are selected from the extracted rules using a heuristic rule sorting criterion to design an accurate and compact classifier [1]-[4]. It is also possible to use global optimization techniques such as genetic algorithms [13], [14] for rule pruning instead of a heuristic rule sorting criterion. Genetic rule selection was first formulated as a single-objective combinatorial optimization problem to design an accurate and compact fuzzy rule-based classifier [15]. A standard single-objective genetic algorithm (SOGA) was used to find a single optimal rule set with respect to a weighted sum fitness function defined by the two objectives: the number of correctly classified training patterns and the number of selected fuzzy rules. This formulation was generalized in [16] where a multiobjective genetic algorithm (MOGA) was used to search for Pareto-optimal fuzzy rule sets with respect to these two objectives. The two-objective formulation in [16] was further generalized to the case of three objectives by taking into account the number of antecedent conditions in each fuzzy rule [17], [18]. The sum of the number of antecedent conditions over selected fuzzy rules was used as an additional complexity measure. An MOGA was used to search for Pareto-optimal rule sets with respect to the three objectives. Multiobjective genetic rule selection was also used to design non-fuzzy rule-based classifiers in [19], [20].

In this paper, we examine the relation between Pareto-optimal fuzzy rules and Pareto-optimal rule sets. First we explain fuzzy rule-based classification, multiobjective fuzzy data mining and multiobjective fuzzy rule selection in Section II. Next we examine whether Pareto-optimal fuzzy rules are included in Pareto-optimal rule sets through computational experiments on some benchmark classification problems in the UCI database in Section III. Then we examine the accuracy of selected fuzzy rules when we use only Pareto-
optimal fuzzy rules as candidate rules in multiobjective genetic rule selection in Section IV. Finally we conclude this paper in Section V.

II. TWO MULTIOBJECTIVE FORMULATIONS

A. Classification problems

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

B. Fuzzy rules

For our $n$-dimensional pattern classification problem, we use fuzzy rules of the following form [21]:

Rule $R_q$: If $x_1$ is $A_{q1}$ and ... and $x_n$ is $A_{qn}$ then Class $C_q$ with $CF_q$,

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

where $R_q$ is the label of the $q$-th fuzzy rule, $\mathbf{x} = (x_1, ..., x_n)$ is an $n$-dimensional pattern vector, $A_{qi}$ is an antecedent fuzzy set, $C_q$ is a class label, and $CF_q$ is a rule weight (i.e., certainty grade). We also denote the fuzzy rule $R_q$ in (1) as $A_q \Rightarrow \text{Class } C_q$. The rule weight $CF_q$ has a large effect on the accuracy of fuzzy rule-based classification systems as shown in [22], [23]. For other types of fuzzy rules for pattern classification problems, see [24]-[26].

Since we usually have no a priori information about an appropriate granularity (i.e., the number of fuzzy sets) of fuzzy discretization for each attribute, we simultaneously use multiple fuzzy partitions with different granularities as shown in Fig. 1. In addition to the 14 fuzzy sets in Fig. 1, we also use the domain interval $[0, 1]$ itself as an antecedent fuzzy set in order to represent a don’t care condition. Thus we have the 15 possible antecedent fuzzy sets as $A_{qi}$ for each attribute.

![Fig. 1. Four fuzzy partitions used in our computational experiments.](image)

C. Fuzzy rule generation

Since we have the 15 antecedent fuzzy sets for each attribute of our $n$-dimensional pattern classification problem, the total number of combinations of the antecedent fuzzy sets is $15^n$. Each combination is used in the antecedent part of the fuzzy rule in (1). Thus the total number of possible fuzzy rules is also $15^n$. The consequent class $C_q$ and the rule weight $CF_q$ of each fuzzy rule $R_q$ are specified from the given training patterns in the following heuristic manner.

First we calculate the compatibility grade of each pattern $\mathbf{x}_p$ with the antecedent part $A_q$ of the fuzzy rule $R_q$ using the product operation as

$$ \mu_{A_q}(\mathbf{x}_p) = \mu_{A_{q1}}(x_{p1}) \cdot ... \cdot \mu_{A_{qn}}(x_{pn}), \quad (2) $$

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

Next the confidence of the fuzzy rule $A_q \Rightarrow \text{Class } h$ is calculated for each class $h$ as follows [26]-[28]:

$$ c(A_q \Rightarrow \text{Class } h) = \frac{\sum_{p=1}^{m} \mu_{A_q}(\mathbf{x}_p)}{\sum_{p=1}^{m} \mu_{A_q}(\mathbf{x}_p)}. \quad (3) $$

The consequent class $C_q$ is specified by identifying the class with the maximum confidence:

$$ c(A_q \Rightarrow \text{Class } C_q) = \max_{h=1,2,\ldots,M} \{c(A_q \Rightarrow \text{Class } h)\}. \quad (4) $$

The consequent class $C_q$ can be viewed as the dominant class in the fuzzy subspace defined by the antecedent part $A_q$. When there is no pattern in the fuzzy subspace defined by $A_q$, we do not generate any fuzzy rules with $A_q$ in the antecedent part. This specification method of the consequent class of fuzzy rules has been used in many studies since [21].

Different specifications of the rule weight $CF_q$ have been proposed and examined. We use the following specification because good results were reported by this specification [23]:

$$ CF_q = c(A_q \Rightarrow \text{Class } C_q) - \sum_{h=1}^{M} \sum_{h \neq C_q} c(A_q \Rightarrow \text{Class } h). \quad (5) $$

D. Rule discovery criteria

Using the above-mentioned procedure, we can generate a large number of fuzzy rules by specifying the consequent class and the rule weight for each of the $15^n$ combinations of the antecedent fuzzy sets. It is, however, very difficult for human users to handle such a large number of generated fuzzy rules. It is also very difficult to intuitively understand long fuzzy rules with many antecedent conditions. Thus we generate short fuzzy rules with a few antecedent conditions. It should be noted that don’t care conditions can be omitted from fuzzy rules. So the rule length means the number of antecedent conditions excluding don’t care conditions. We examine only short fuzzy rules of length $L_{max}$ or less (e.g., $L_{max} = 3$). This restriction is to find a compact set of fuzzy rules with high interpretability.

Among short fuzzy rules, we only extract fuzzy rules that satisfy both the minimum confidence and support. In the field
of data mining, these two rule extraction criteria have been frequently used [1]-[5]. In the same manner as the fuzzy version of confidence in (3), the support of the fuzzy rule \( A_q \Rightarrow \text{Class} \ h \) is calculated as follows [26]-[28]:

\[
    s(A_q \Rightarrow \text{Class} \ h) = \frac{\sum_{x_p \in \text{Class} \ h} \mu_{A_q}(x_p)}{m}.
\]  

(6)

E. Multiobjective fuzzy rule mining

The use of NSGA-II [29], which is a well-known and frequently-used multiobjective genetic algorithm (MOGA), was proposed by de la Iglesia et al. [7]-[9] in the field of classification rule mining. They applied NSGA-II to the following two-objective rule discovery problem for partial classification.

Maximize \{Confidence(R), Coverage(R)\},

(7)

where \( R \) means a classification rule. It should be noted in (7) that the coverage maximization is the same as the support maximization since the consequent class is always fixed in partial classification. The use of a rule dissimilarity measure between classification rules instead of a crowding measure in NSGA-II was examined in [8] in order to search for a set of Pareto-optimal rules with a large diversity. Pareto-dominance relation in NSGA-II was modified in [10] in order to search for not only Pareto-optimal rules but also near Pareto-optimal rules. The two-objective formulation in (7) is used to define Pareto-optimal fuzzy rules in this paper.

F. Multiobjective fuzzy rule selection

Let \( S \) be a set of fuzzy rules of the form in (1). That is, \( S \) is a fuzzy rule-based classifier. A new pattern \( x_p \) is classified by a single winner rule \( R_w \), which is chosen from the rule set \( S \) as follows:

\[
    \mu_{A_w}(x_p) \cdot CF_w = \max \{ \mu_{A_q}(x_p) \cdot CF_q \mid R_q \in S \}.
\]  

(8)

When multiple fuzzy rules with different consequent classes have the same maximum value, \( x_p \) is randomly assigned to one of those classes. On the other hand, the classification of \( x_p \) is rejected when no rules are compatible with \( x_p \) (i.e., when the maximum value is zero in (8)). Such a pattern is viewed as being unclassifiable by \( S \).

The above-mentioned random tiebreak is not used in the rule selection phase in which the combination of fuzzy rules is optimized. The random tiebreak is used only when the accuracy of selected rule sets (i.e., obtained fuzzy rule-based classifiers) is evaluated after rule selection.

As in our former studies [17]-[19], we use the following three objectives in multiobjective genetic rule selection:

\[
    f_1(S) : \text{The number of correctly classified patterns by } S,
\]

\[
    f_2(S) : \text{The number of selected fuzzy rules in } S,
\]

\[
    f_3(S) : \text{The total number of antecedent conditions in } S \text{ (i.e., the total rule length in } S)\).
\]

The first objective is maximized while the second and third ones are minimized. That is, our three-objective rule selection problem is written as follows:

Maximize \( f_1(S) \), and minimize \{ \( f_2(S) \), \( f_3(S) \) \}.

(9)

When the first objective is to be evaluated in multiobjective genetic rule selection, we use the single winner-based method without the random tiebreak. We apply NSGA-II to the three-objective rule selection problem in (9). For details of NSGA-II, see [29], [30]. For the implementation of three-objective genetic rule selection, see [17]-[20].

III. EXAMINATION OF SELECTED FUZZY RULES

A. Settings of computational experiments

We used the five data sets in Table 1 from the UCI database. Incomplete patterns with missing values were not used. All attribute values were handled as real numbers in the unit interval [0, 1]. We simultaneously used the four different fuzzy partitions with two, three, four, and five membership functions in Fig. 1. That is, we used 14 fuzzy sets and don’t care as antecedent fuzzy sets for each attribute.

<table>
<thead>
<tr>
<th>TABLE I DATA SETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set</td>
</tr>
<tr>
<td>Breast W</td>
</tr>
<tr>
<td>Glass</td>
</tr>
<tr>
<td>Heart C</td>
</tr>
<tr>
<td>Iris</td>
</tr>
<tr>
<td>Wine</td>
</tr>
</tbody>
</table>

* Incomplete patterns with missing values are not included.

We divided each data set into two subsets of the same size: training data and test data. Using training data, first we extracted fuzzy rules satisfying the following conditions:

Minimum confidence: 0.6,
Minimum support: 0.04 (for the wine data set),
0.01 (for the other data sets).

The maximum rule length was specified as three. All the extracted fuzzy rules were used in multiobjective genetic rule selection as candidate rules.

Then we applied NSGA-II to the generated candidate rules to search for Pareto-optimal rule sets with respect to the three objectives in (9) using the following parameter specifications:

Population size: 200 strings,
Crossover probability: 0.9 (uniform crossover),
Mutation probability: 0.05 (1 \rightarrow 0),
\( 1/N (0 \rightarrow 1, N: \text{string length}) \),
Termination condition: 1000 generations.

In our multiobjective genetic rule selection, the string length \( N \) is the same as the number of the generated candidate rules.
B. Experimental results

**Wisconsin Breast Cancer Data Set:** First we randomly divided the data set into 342 and 341 patterns for training and testing, respectively. Next we extracted 78607 candidate fuzzy rules from the 342 training patterns. Then we applied NSGA-II to the candidate rules for multiobjective genetic rule selection. From its single run, 18 non-dominated rule sets were obtained. Finally each of the obtained rule sets was evaluated for the training patterns and the test patterns.

Training data accuracy and test data accuracy of each rule set are shown in Fig. 2 (a) and Fig. 2 (b), respectively. We can observe a clear accuracy-complexity tradeoff relation in Fig. 2 (a): The classification rate increases with the increase in the number of fuzzy rules. This means that we can not simultaneously realize the accuracy maximization and the complexity minimization. Since we used not only the number of fuzzy rules but also the total number of antecedent conditions as complexity measures in multiobjective genetic rule selection, different rule sets with the same number of fuzzy rules were obtained in Fig. 2 (a). Rule sets marked by A and B are examined later. Some of the obtained rule sets (i.e., a rule set with a single rule) are not shown because their classification rates are out of the range of the vertical axis of each plot in Fig. 2.

![Fig. 2. Obtained fuzzy rule sets (Breast W).](image)

We examined fuzzy rules included in each of the obtained rule sets. That is, we examined whether selected fuzzy rules are Pareto-optimal or not. In Fig. 3, we show all the selected fuzzy rules together with the candidate rules. The upper-right bound of the cloud of the candidate rules is the Pareto front in terms of confidence and coverage maximization (see Fig. 14). Fuzzy rules marked by open circles in Fig. 3 were included in at least one of the obtained rule sets. In Fig. 3, some selected rules are Pareto-optimal or near Pareto-optimal (i.e., they are very close to the upper-right bound of the cloud of the candidate rules in each plot of Fig. 3). Other selected rules, however, are far from Pareto-optimal rules. In Fig. 4 and Fig. 5, we show selected fuzzy rules in the rule sets A and B in Fig. 2, respectively. As shown in Fig. 4, small rule sets usually consist of only Pareto-optimal or near Pareto-optimal fuzzy rules. This is because each rule in small rule sets should classify many patterns. On the contrary, large rule sets often include fuzzy rules far from Pareto-optimal rules as shown in Fig. 5.

![Fig. 3. Selected rules in Fig. 2 (Breast W).](image)

**Glass Data Set:** Experimental results are shown in Fig. 6. Multiobjective genetic rule selection found 24 rule sets (some of them are not shown in Fig. 6 due to their low accuracy). We can observe a clear accuracy-complexity tradeoff relation in Fig. 6 (a). In Fig. 7, we show selected fuzzy rules in the obtained 24 rule sets. Due to the page limitation, we only show selected rules with Class 1 and Class 2 consequents in Fig. 7 (a) and Fig. 7 (b), respectively. As in Fig. 3, some
selected rules are Pareto-optimal or near Pareto-optimal while others are far from Pareto-optimal rules.

![Graph](image1)

Fig. 6. Obtained fuzzy rule sets (Glass).

![Graph](image2)

Fig. 7. Selected rules in Fig. 6 (Glass).

**Cleveland Heart Disease Data Set:** Experimental results are shown in Fig. 8. Multiobjective genetic rule selection found 33 rule sets. We can observe a clear accuracy-complexity tradeoff relation in Fig. 8 (a). On the other hand, we can observe a slight overfitting to training data in Fig. 8 (b). That is, the test data accuracy slightly decreases with the increase in the number of fuzzy rules in the range with more than four fuzzy rules. Selected rules are shown for Class 1 and Class 2 in Fig. 9 (a) and Fig. 9 (b), respectively. Some selected rules are Pareto-optimal or near Pareto-optimal while others are far from Pareto-optimal rules.

**Iris Data Set:** Experimental results are shown in Fig. 10. All the training patterns were correctly classified by four fuzzy rules in Fig. 10 (a). Selected rules are shown in Fig. 11. As we can see from Fig. 11, almost all the selected rules are Pareto-optimal or near Pareto-optimal. This may be because all the selected rule sets are small in Fig. 10 with four or less fuzzy rules.

**Wine Data Set:** Experimental results are shown in Fig. 12 and Fig. 13. All the training patterns were correctly classified by four fuzzy rules in Fig. 12 (a). As in Fig. 11 for the iris data set, almost all the selected rules are Pareto-optimal or near Pareto-optimal in Fig. 13 for the wine data set.
IV. PARETO-OPTIMAL CANDIDATE FUZZY RULES

A. Settings of computational experiments

For comparison, we performed the same computational experiments as in Section III using only Pareto-optimal rules as candidate rules in multiobjective genetic rule selection. In Fig. 14, we show Pareto-optimal rules for the Wisconsin breast cancer data set. Pareto-optimal rules with respect to confidence maximization and coverage maximization are marked by open circles in Fig. 14. These rules were used as candidate rules in multiobjective genetic rule selection in this section. Except for the choice of the candidate rules, we used the same parameter specifications as in Section III.

B. Experimental results

Wisconsin Breast Cancer Data Set: Experimental results are shown in Fig. 15. From the comparison between Fig. 2 and Fig. 15, we can see that large rule sets with many rules were not obtained when we used only Pareto-optimal rules as candidate rules. This observation is consistent with Fig. 5 where some selected rules in a large rule set were far from Pareto-optimal rules. Whereas high training data accuracy was not obtained in Fig. 15 (a) in comparison with Fig. 2 (a), good test data accuracy was obtained in Fig. 15 (b). That is, the test data accuracy was not severely degraded in Fig. 15 (b) from Fig. 2 (b). Almost the same observations were obtained from computational experiments on the other four data sets as shown later.

Glass Data Set: Experimental results are shown in Fig. 16. Whereas the training data accuracy was degraded from Fig. 6 (a) by about 10%, almost the same test data accuracy was obtained in Fig. 16 (b) as in Fig. 6 (b).

Cleveland Heart Disease Data Set: Experimental results are shown in Fig. 17. From the comparison between Fig. 8 and Fig. 17, we can see that almost the same test data accuracy was obtained while the training data accuracy was degraded by about 10%.
V. CONCLUDING REMARKS

In this paper, we examined the relation between Pareto-optimal rules and Pareto-optimal rule sets in the design of accurate and compact fuzzy rule-based classifiers through computational experiments using multiobjective genetic rule selection. First we checked whether Pareto-optimal rules were selected by multiobjective genetic fuzzy rule selection. Computational experiments showed that some selected rules were Pareto-optimal or near Pareto-optimal while others are far from Pareto-optimal rules. Experimental results also showed that large rule sets with many rules tended to include fuzzy rules far from Pareto-optimal rules. Next we examined the classification performance of obtained rule sets when we used Pareto-optimal rules as candidate rules. Experimental results showed that the use of Pareto-optimal rules as candidate rules degraded training data accuracy while it did not degrade test data accuracy in many cases.

Since almost all selected rules were Pareto-optimal or near Pareto-optimal in small rule sets with high interpretability, it may be a promising strategy to use Pareto-optimal and near Pareto-optimal rules as candidate rules (instead of all rules satisfying the minimum support and confidence) to efficiently find accurate and interpretable fuzzy rule-based classifiers by multiobjective genetic rule selection.

ACKNOWLEDGEMENT

This work was partially supported by Grant-in-Aid for Scientific Research on Priority Areas (18049065) and for Scientific Research (B) (17300075).

REFERENCES


---

**Iris Data Set**: Experimental results are shown in Fig. 18. A 100% training data accuracy was not obtained in Fig. 18 (a) while all the training patterns were correctly classified by four fuzzy rules in Fig. 10. The test data accuracy, however, was not degraded (compare Fig. 18 (b) with Fig. 10 (b)).

**Wine Data Set**: Experimental results are shown in Fig. 19. From the comparison between Fig. 12 and Fig. 19, we can see that the use of Pareto-optimal rules as candidate rules degraded the test data accuracy as well as the training data accuracy. Such a clear degrade in the test data accuracy was observed only for the wine data among the five data sets.


