Finding Simple Fuzzy Classification Systems with High Interpretability Through Multiobjective Rule Selection

Hisao Ishibuchi, Yusuke Nojima, and Isao Kuwajima

Department of Computer Science and Intelligent Systems, Graduate School of Engineering
Osaka Prefecture University
1-1 Gakuen-cho, Sakai, Osaka 599-8531, Japan
hisaoi@cs.osakafu-u.ac.jp, nojima@cs.osakafu-u.ac.jp
kuwajima@ci.cs.osakafu-u.ac.jp
http://www.ie.osakafu-u.ac.jp/~hisaoi/

Abstract. In this paper, we demonstrate that simple fuzzy rule-based classification systems with high interpretability are obtained through multiobjective genetic rule selection. In our approach, first a prespecified number of candidate fuzzy rules are extracted from numerical data in a heuristic manner using rule evaluation criteria. Then multiobjective genetic rule selection is applied to the extracted candidate fuzzy rules to find a number of non-dominated rule sets with respect to the classification accuracy and the complexity. The obtained non-dominated rule sets form an accuracy-complexity tradeoff surface. The performance of each non-dominated rule set is evaluated in terms of its classification accuracy and its complexity. Computational experiments show that our approach finds simple fuzzy rules with high interpretability for some benchmark data sets in the UC Irvine machine learning repository.

1 Introduction

Evolutionary multiobjective optimization (EMO) is one of the most active research areas in the field of evolutionary computation [2], [3], [7]. The main advantage of EMO algorithms over classical approaches is that a number of non-dominated solutions can be obtained by a single run of EMO algorithms. EMO algorithms have been successfully applied to various application areas [2], [3], [7]. In the application to fuzzy logic, EMO algorithms have been used to find accurate, transparent and compact fuzzy rule-based systems [6], [16]-[18]. That is, EMO algorithms have been used to maximize the accuracy of fuzzy rule-based systems and minimize their complexity.

In this paper, we clearly demonstrate that simple fuzzy rule-based classification systems with high interpretability can be obtained by multiobjective fuzzy rule selection. We also demonstrate that a number of non-dominated fuzzy rule-based classification systems along the accuracy-complexity tradeoff surface can be obtained by a single run of an EMO-based fuzzy rule selection algorithm. Fuzzy rule selection for classification problems was first formulated as a single-objective combinatorial optimization problem [13], [14]. A standard genetic algorithm was used to optimize a weighted sum fitness function, which was defined by the number of correctly

B. Gabrys, R.J. Howlett, and L.C. Jain (Eds.): KES 2006, Part II, LNAI 4252, pp. 86 – 93, 2006.
© Springer-Verlag Berlin Heidelberg 2006

classified training patterns and the number of fuzzy rules. A two-objective combinatorial optimization problem was formulated in [10] as an extension of the single-objective formulation. An EMO algorithm was used to find a number of non-dominated fuzzy rule-based classification systems with respect to the two objectives: maximization of the number of correctly classified training patterns and minimization of the number of selected fuzzy rules. The two-objective formulation was further extended in [11], [15] to a three-objective combinatorial optimization problem by introducing an additional objective: minimization of the total number of antecedent conditions (i.e., minimization of the total rule length). A number of non-dominated fuzzy rule-based systems with respect to the three objectives were found by an EMO algorithm.

This paper is organized as follows. First we explain an outline of multiobjective fuzzy rule selection in Section 2. Next we explain heuristic rule extraction for extracting candidate rules in Section 3. Then we show experimental results on some benchmark data sets in the UC Irvine machine learning repository in Section 4. Finally we conclude this paper in Section 5.

2 Multiobjective Fuzzy Rule Selection

Let us assume that we have N fuzzy rules as candidate rules for multiobjective fuzzy rule selection. We denote a subset of those candidate rules by S. The accuracy of the rule set S is measured by the error rate on the given training patterns. We use a single winner rule-based method [12] to classify each training pattern by S. The single winner rule for a training pattern has the maximum product of the rule weight and the compatibility grade with that pattern. We include the rejection rate into the error rate (i.e., training patterns with no compatible fuzzy rules in S are counted among errors).

On the other hand, we measure the complexity of the rule set S by the number of fuzzy rules in S and the total number of antecedent conditions in S. Thus our multiobjective fuzzy rule selection problem is formulated as follows:

$$\text{Minimize } f_1(S),\ f_2(S),\ f_3(S), \tag{1}$$

where $f_1(S)$ is the error rate on training patterns, $f_2(S)$ is the number of fuzzy rules, and $f_3(S)$ is the total number of antecedent conditions. It should be noted that each rule has a different number of antecedent conditions. This is because we use *don't care* as a special antecedent fuzzy set, which is not counted as antecedent conditions. That is, the third objective is the number of antecedent conditions excluding *don't care* conditions. The third objective can be also viewed as the total rule length since the number of antecedent conditions is often referred to as the rule length.

Any subset S of the N candidate fuzzy rules can be represented by a binary string of length N as $S = s_1 s_2 \dots s_N$ where $s_j = 1$ and $s_j = 0$ mean that the j-th rule is included in S and excluded from S, respectively. Such a binary string is handled as an individual in multiobjective fuzzy rule selection. Since individuals are represented by binary strings, we can apply almost all EMO algorithms with standard genetic operations to our multiobjective fuzzy rule selection problem in (1). In this paper, we use the NSGA-II algorithm [8] due to its popularity, high performance and simplicity.

In the application of NSGA-II to multiobjective fuzzy rule selection, we use two heuristic tricks to efficiently find small rule sets with high accuracy. One trick is biased mutation where a larger probability is assigned to the mutation from 1 to 0 than that from 0 to 1. The other trick is the removal of unnecessary rules, which is a kind of local search. Since we use the single winner rule-based method for the classification of each pattern by the rule set S, some rules in S may be chosen as winner rules for no training patterns. By removing those rules from S, we can improve the second and third objectives without degrading the first objective. The removal of unnecessary rules is performed after the first objective is calculated and before the second and third objectives are calculated. NSGA-II with these two tricks is used to find non-dominated rule sets of the multiobjective fuzzy rule selection problem in (1).

Here we briefly explain some basic concepts in multiobjective optimization. Let us consider the following k-objective minimization problem:

$$\text{Minimize } \mathbf{z} = (f_1(\mathbf{y}),\ f_2(\mathbf{y}),\ ...,\ f_k(\mathbf{y})) \text{ subject to } \mathbf{y} \in \mathbf{Y}, \tag{2}$$

where $\mathbf{z}$ is the objective vector, $f_i(\mathbf{y})$ is the i-th objective to be minimized, $\mathbf{y}$ is the decision vector, and $\mathbf{Y}$ is the feasible region in the decision space.

Let $\mathbf{a}$ and $\mathbf{b}$ be two feasible solutions of the k-objective minimization problem in (2). If the following condition holds, $\mathbf{a}$ can be viewed as being better than $\mathbf{b}$:

$$\forall\, i,\ f_i(\mathbf{a}) \le f_i(\mathbf{b}) \ \text{ and } \ \exists\, j,\ f_j(\mathbf{a}) < f_j(\mathbf{b}). \tag{3}$$

In this case, we say that $\mathbf{a}$ dominates $\mathbf{b}$ (equivalently $\mathbf{b}$ is dominated by $\mathbf{a}$).

When $\mathbf{b}$ is not dominated by any other feasible solutions (i.e., when there exists no feasible solution $\mathbf{a}$ that dominates $\mathbf{b}$), the solution $\mathbf{b}$ is referred to as a Pareto-optimal solution of the k-objective minimization problem in (2). The set of all Pareto-optimal solutions forms the tradeoff surface in the objective space. Various EMO algorithms have been proposed to efficiently search for Pareto-optimal solutions [2], [3], [7]. Since it is very difficult to find the true Pareto-optimal solutions of a large-scale multiobjective optimization problem, non-dominated solutions among the examined ones during the execution of EMO algorithms are usually presented as a final solution set.

3 Heuristic Fuzzy Rule Extraction

Let us assume that we have m training patterns $\mathbf{x}_p = (x_{p1}, x_{p2}, ..., x_{pn})$, $p = 1, 2, ..., m$ 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].

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

$$\text{Rule } R_q : \text{If } x_1 \text{ is } A_{q1} \text{ and ... and } x_n \text{ is } A_{qn} \text{ then Class } C_q \text{ with } CF_q, \tag{4}$$

where R_q is the label of the q-th fuzzy rule, $\mathbf{x} = (x_1, x_2, ..., x_n)$ 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 denote the fuzzy rule R_q in (4) as "$\mathbf{A}_q \Rightarrow$ Class C_q" where $\mathbf{A}_q = (A_{q1}, A_{q2}, ..., A_{qn})$.

Since we usually have no *a priori* information about an appropriate granularity of the fuzzy discretization for each attribute, we simultaneously use multiple fuzzy partitions with different granularities to extract candidate fuzzy rules. In computational experiments, we use four homogeneous fuzzy partitions with triangular fuzzy sets in Fig. 1. In addition to the 14 fuzzy sets in Fig. 1, we also use the domain interval [0, 1] as an antecedent fuzzy set in order to represent a *don't care* condition. That is, we use the 15 antecedent fuzzy sets for each attribute in computational experiments.

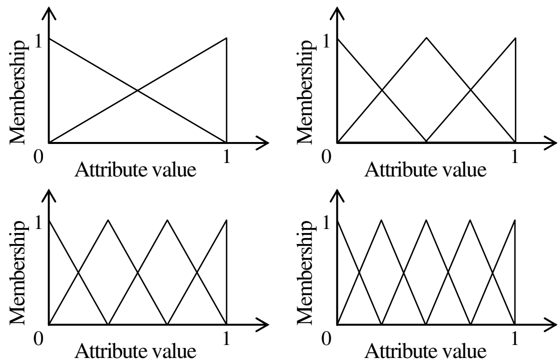

Fig. 1. Antecedent fuzzy sets used in computational experiments

Since we use 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 (4). 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 can be specified from compatible training patterns in a heuristic manner (for details, see Ishibuchi et al. [12]). That is, 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 examine only short fuzzy rules of length $L_{\max}$ or less (e.g., $L_{\max} = 3$). This restriction on the rule length (i.e., the number of antecedent conditions) is to find rule sets of simple fuzzy rules with high interpretability.

Among short fuzzy rules, we generate a prespecified number of candidate fuzzy rules using heuristic rule evaluation criteria. In the field of data mining, two rule evaluation criteria (i.e., confidence and support) have been often used [1], [4], [5]. The fuzzy version of the confidence is defined as follows [12]:

$$c(\mathbf{A}_q \Rightarrow \text{Class } h) = \sum_{\mathbf{x}_p \in \text{Class } h} \mu_{\mathbf{A}_q}(\mathbf{x}_p) \Big/ \sum_{p=1}^{m} \mu_{\mathbf{A}_q}(\mathbf{x}_p)\,, \tag{5}$$

where $\mu_{\mathbf{A}q}(\mathbf{x}_p)$ is the compatibility grade of each training pattern $\mathbf{x}_p$ with the antecedent part $\mathbf{A}_q$ of the fuzzy rules $\mathbf{A}_q \Rightarrow$ Class C_q in (4), which is defined as follows:

$$\mu_{\mathbf{A}_q}(\mathbf{x}_p) = \mu_{A_{q1}}(x_{p1}) \cdot \ ... \ \cdot \mu_{A_{qn}}(x_{pn})\,. \tag{6}$$

In the same manner, the support is defined as follows [12]:

$$s(\mathbf{A}_q \Rightarrow \text{Class } h) = \sum_{\mathbf{x}_p \in \text{Class } h} \mu_{\mathbf{A}_q}(\mathbf{x}_p) \Big/ m\,. \tag{7}$$

A prespecified number of candidate fuzzy rules are extracted using the following ranking mechanisms of fuzzy rules (Of course, we can use other methods such as the SLAVE criterion [9]):

Support criterion with the minimum confidence level: Each fuzzy rule is evaluated based on its support when its confidence is larger than or equal to the prespecified minimum confidence level. Under this criterion, we never extract unqualified rules whose confidence is smaller than the minimum confidence level. Various values of the minimum confidence level are examined in computational experiments.

Confidence criterion with the minimum support level: Each fuzzy rule is evaluated based on its confidence when its support is larger than or equal to the prespecified minimum support level. Under this criterion, we never extract unqualified rules whose support is smaller than the minimum support level. Various values of the minimum support level are examined in computational experiments.

4 Computational Experiments

In the first stage of our multiobjective rule selection method, a prespecified number of candidate fuzzy rules are extracted. We extract 300 candidate fuzzy rules for each class using the above-mentioned two ranking mechanisms. Experimental results on some benchmark data sets in the UC Irvine machine learning repository are summarized in Table 1 where the classification rates on training patterns of 300*M* candidate fuzzy rules are shown (*M*: the number of classes in each data set). In the last column, we extract 30 candidate fuzzy rules using the ten specifications in the other columns and use all of them (i.e., 300 candidate fuzzy rules for each class in total). Bold face shows the best result for each data set. Bad results, which are more than 10% worse than the best result, are indicated by underlines. From Table 1, we can see that the choice of an appropriate specification in the rule ranking mechanisms is problem-specific. Since no specification is good for all the seven data sets, we use 30 candidate rules extracted by each of the ten specifications (i.e., 300 rules for each class in the last column of Table 1: 300*M* rules for each data set).

Then we apply NSGA-II to the 300*M* candidate fuzzy rules for each data set to search for non-dominated rule sets. Fig. 2 shows five rule sets obtained by a single run of NSGA-II for the iris data. It should be noted that the three plots show the same five rule sets using a different horizontal axis. We can observe in each plot of Fig. 2 the accuracy-complexity tradeoff with respect to a different complexity measure. One rule set with a 2.67% error rate (i.e., the simplest rule set in Fig. 2) and the corresponding classification boundary are shown in Fig. 3. We can see that the rule set with the three fuzzy rules in Fig. 3 has high linguistic interpretability. For example, the first rule is linguistically interpreted as "If x_4 is *very small* then Class 1." We can also see that the classification boundary in Fig. 3 is intuitively acceptable.

Table 1. Classification rates on training patterns of candidate fuzzy rules

Data set	Support with minimum conf.					Confidence with minimum sup.					Mixed
	0.6	0.7	0.8	0.9	1.0	0.01	0.02	0.05	0.10	0.15	
Iris	**96.00**	**96.00**	**96.00**	**96.00**	88.00	92.67	94.00	**96.00**	**96.00**	95.33	**96.00**
Breast W	95.90	95.90	95.90	95.90	82.43	90.48	94.58	**96.78**	**96.78**	96.24	96.34
Diabetes	69.40	69.92	73.05	**78.26**	14.06	63.28	64.45	77.34	75.52	71.61	70.44
Glass	**69.63**	65.89	56.07	31.78	23.83	55.61	69.16	62.62	64.49	59.44	68.69
Heart C	62.96	64.65	**68.35**	58.92	48.82	54.21	55.62	50.17	59.26	52.46	65.32
Sonar	77.40	78.37	79.33	**90.38**	83.65	81.25	87.98	88.94	79.33	77.88	87.02
Wine	97.19	97.19	96.63	97.19	**98.88**	95.51	96.07	**98.88**	96.63	96.07	96.63

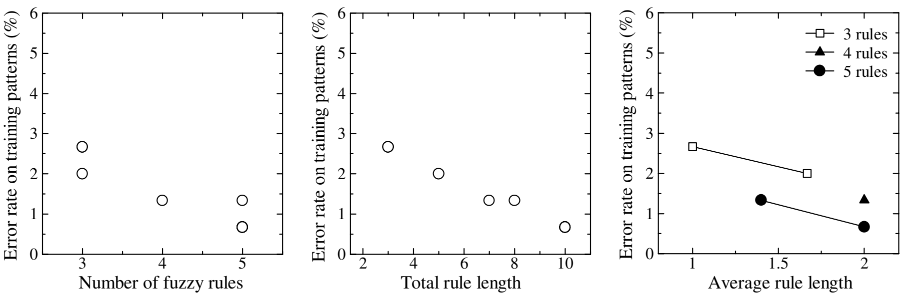

Fig. 2. Non-dominated rule sets by a single run of the NSGA-II algorithm for the iris data

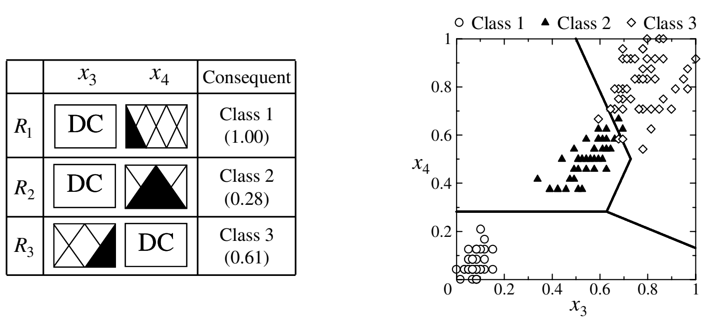

Fig. 3. One rule set obtained for the iris data and the corresponding classification boundary

We also obtain simple rule sets with high interpretability for the other data sets as shown in Fig. 4 where the error rate of each data set is shown as a figure caption.

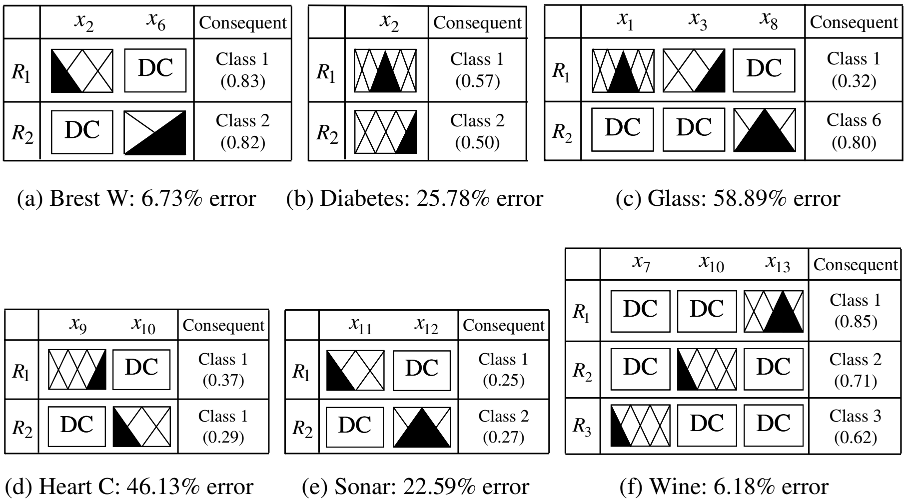

Fig. 4. Example of obtained rule sets for the other data sets

5 Concluding Remarks

In this paper, we demonstrated that simple rule sets with high interpretability can be obtained by multiobjective rule selection for some benchmark data sets in the UC Irvine machine learning repository. Since our approach selects a small number of short fuzzy rules using homogeneous fuzzy partitions, rule sets with high linguistic interpretability are obtained as shown in Fig. 3 and Fig. 4. Whereas all the rule sets in Fig. 3 and Fig. 4 have high interpretability, the classification accuracy of these rule sets is not necessarily high. Especially the rule sets in Fig. 4 (c) for the glass data with six classes and Fig. 4 (d) for the Cleveland heart disease data with five classes have poor classification accuracy. This is because the number of fuzzy rules is less than the number of classes. When emphasis should be placed on the classification accuracy, more complicated rule sets with higher accuracy can be chosen from non-dominated rule sets. For example, our approach found a rule set with 14 fuzzy rules for the glass data. The error rate of this rule set was 17.76% whereas the rule set with two fuzzy rules in Fig. 4 (c) has a 58.89% error rate. Another difficulty of our approach for multi-class problem is that fuzzy rules for all classes are not necessarily included in rule sets. When we need at least one fuzzy rule for each class, an additional constraint condition can be introduced to multiobjective rule selection.

Acknowledgement

This work was partially supported by Japan Society for the Promotion of Science (JSPS) through Grand-in-Aid for Scientific Research (B): KAKENHI (17300075).

References

1. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A. I.: Fast Discovery of Association Rules. In: Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.): *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park (1996) 307-328
2. Coello Coello, C. A., Lamont, G. B.: *Applications of Multi-Objective Evolutionary Algorithms*. World Scientific, Singapore (2004)
3. Coello Coello, C. A., Van Veldhuizen, D. A., Lamont, G. B.: *Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer Academic Publishers, Boston, MA (2002)
4. Coenen, F., Leng, P.: Obtaining Best Parameter Values for Accurate Classification. *Proc. of the 5th IEEE International Conference on Data Mining* (2005) 549-552
5. Coenen, F., Leng, P., and Zhang, L.: Threshold Tuning for Improved Classification Association Rule Mining. *Lecture Notes in Computer Science 3518: Advances in Knowledge Discovery And Data Mining - PAKDD 2005*. Springer, Berlin (2005) 216-225
6. Cordon, O., Jesus, M. J. del, Herrera, F., Magdalena, L., and Villar, P.: A Multiobjective Genetic Learning Process for Joint Feature Selection and Granularity and Contexts Learning in Fuzzy Rule-based Classification Systems. In: Casillas, J., Cordon, O., Herrera, F., Magdalena, L. (eds.): *Interpretability Issues in Fuzzy Modeling*. Springer, Berlin (2003) 79-99
7. Deb, K.: *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Chichester (2001)
8. Deb, K., Pratap, A., Agrawal, S., Meyarivan, T.: A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Trans. on Evolutionary Computation 6*, 2 (2002) 182-197
9. Gonzalez, A., Perez, R.: SLAVE: A Genetic Learning System based on an Iterative Approach. *IEEE Trans. on Fuzzy Systems 7*, 2 (1999) 176-191
10. Ishibuchi, H., Murata, T., Turksen, I. B.: Single-Objective and Two-Objective Genetic Algorithms for Selecting Linguistic Rules for Pattern Classification Problems. *Fuzzy Sets and Systems 89*, 2 (1997) 135-150
11. Ishibuchi, H., Nakashima, T., Murata, T.: Three-Objective Genetics-based Machine Learning for Linguistic Rule Extraction. *Information Sciences 136*, 1-4 (2001) 109-133
12. Ishibuchi, H., Nakashima, T., Nii, M.: *Classification and Modeling with Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining*. Springer, Berlin (2004)
13. Ishibuchi, H., Nozaki, K., Yamamoto, N., Tanaka, H.: Construction of Fuzzy Classification Systems with Rectangular Fuzzy Rules Using Genetic Algorithms. *Fuzzy Sets and Systems 65*, 2/3 (1994) 237-253
14. Ishibuchi, H., Nozaki, K., Yamamoto, N., Tanaka, H.: Selecting Fuzzy If-Then Rules for Classification Problems Using Genetic Algorithms. *IEEE Trans. on Fuzzy Systems 3*, 3 (1995) 260-270
15. Ishibuchi, H., Yamamoto, T.: Fuzzy Rule Selection by Multi-Objective Genetic Local Search Algorithms and Rule Evaluation Measures in Data Mining. *Fuzzy Sets and Systems 141*, 1 (2004) 59-88
16. Jimenez, F., Gomez-Skarmeta, A. F., Sanchez, G., Roubos, H., Babuska, R.: Accurate, Transparent and Compact Fuzzy Models by Multi-Objective Evolutionary Algorithms. In Casillas, J., Cordon, O., Herrera, F., Magdalena, L. (eds.): *Interpretability Issues in Fuzzy Modeling*. Springer, Berlin (2003) 431-451
17. Wang, H., Kwong, S., Jin, Y., Wei, W., Man, K. F.: Multi-Objective Hierarchical Genetic Algorithm for Interpretable Fuzzy Rule-based Knowledge Extraction. *Fuzzy Sets and Systems 149*, 1 (2005) 149-186
18. Wang, H., Kwong, S., Jin, Y., Wei, W., Man, K. F.: Agent-based Evolutionary Approach for Interpretable Rule-based Knowledge Extraction. *IEEE Trans. on Systems, Man, and Cybernetics - Part C: Applications and Reviews 35*, 2 (2005) 143-155