Effects of Diversity Measures on the Design of Ensemble Classifiers by Multiobjective Genetic Fuzzy Rule Selection with a Multi-classifier Coding Scheme

Yusuke Nojima and Hisao Ishibuchi

Dept. of Computer Science and Intelligent Systems, Graduate School of Engeering,
Osaka Prefecture University
1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531, Japan
{nojima,hisaoi}@cs.osakafu-u.ac.jp

Abstract. We have already proposed multiobjective genetic fuzzy rule selection with a multi-classifier coding scheme for the design of ensemble classifiers. An entropy-based diversity measure was used as an objective to be maximized for increasing the diversity of base classifiers in an ensemble. In this paper, we examine the use of other diversity measures in the design of ensemble classifiers. Experimental results show that the choice of a diversity measure has a large effect on the performance of designed ensemble classifiers.

Keywords: Genetic Fuzzy Systems, Ensemble Classifiers, Diversity Measures.

1 Introduction

The use of multiple classifiers as an ensemble is a promising approach to the design of reliable classifiers [1, 2]. In the design of ensemble classifiers, it is essential to generate a number of base classifiers with a large diversity. Some studies [1, 2] rely on diversity maintenance mechanisms to generate base classifiers while others [3, 4] use heuristic measures to evaluate the diversity of base classifiers. Such a heuristic measure can be incorporated into evolutionary approaches for generating a number of base classifiers. Evolutionary multiobjective optimization (EMO) algorithms have been frequently used in the design of ensemble classifiers where the accuracy and the diversity of base classifiers are simultaneously optimized [5-9].

In our former studies [8, 9], we have proposed multiobjective genetic fuzzy rule selection with a multi-classifier coding scheme for the design of ensemble classifiers. Genetic fuzzy rule selection was originally proposed in [10] where a small number of fuzzy rules were selected from a large number of candidate rules to generate a fuzzy rule-based classifier. This method was extended to multiobjective formulations in [11, 12]. A binary string was used to represent a fuzzy rule-based classifier (i.e., a subset of candidate rules) in these studies [10-12]. The use of binary strings was generalized by introducing a multi-classifier coding scheme in our former studies [8, 9] where an integer string was used to represent an ensemble classifier (i.e., an ensemble of disjoint subsets of candidate rules). Each ensemble was evaluated by its classification accuracy

E. Corchado, A. Abraham, and W. Pedrycz (Eds.): HAIS 2008, LNAI 5271, pp. 755–763, 2008.

and the entropy of outputs from its base classifiers. In this paper, we examine the use of various diversity measures [4] in our evolutionary ensemble method with the multi-classifier coding scheme.

In this paper, we first explain our EMO-based approach to the design of ensemble classifiers. Next we examine the use of various diversity measures in our approach in Section 3. Finally we conclude this paper in Section 4.

2 Fuzzy Ensemble Classifiers by Genetic Fuzzy Rule Selection

Our ensemble design method consists of two stages: candidate fuzzy rule extraction and ensemble classifier optimization. In the first stage (i.e., candidate fuzzy rule extraction), a prespecified number of fuzzy rules are extracted in a heuristic manner.

Let us assume that we have m training patterns $\mathbf{x}_p = (x_{p1}, ..., x_{pn})$, $p = 1, 2, ..., m$ from M classes where x_{pi} is the attribute value of the pth training pattern for the ith attribute ($i = 1, 2, ..., n$). We also assume that all attribute values have already been normalized into real numbers in the unit interval [0, 1]. For our n-dimensional M-class problem, we use fuzzy rules of the following form:

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

where R_q is the label of the qth rule, A_{qi} is an antecedent fuzzy set, C_q is a class label, and CF_q is a rule weight. As antecedent fuzzy sets, we use the 14 fuzzy sets in Fig. 1 and "*don't care*" represented by the unit interval [0, 1]. Thus the total number of combinations of the n antecedent fuzzy sets in (1) is 15^n. Among these possible rules, we examine only short fuzzy rules with a small number of antecedent conditions (i.e., short fuzzy rules with many *don't care* conditions) to generate candidate rules.

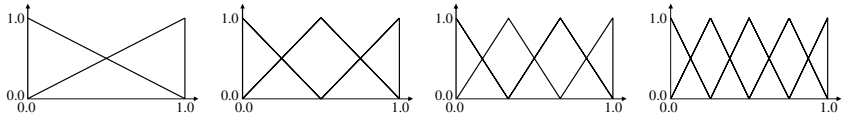

Fig. 1. Fourteen antecedent fuzzy sets used in this paper

For determining the consequent class C_q and the rule weight CF_q, first the confidence of the fuzzy rule "$\mathbf{A}_q \Rightarrow$ Class h" is calculated for each class h as follows:

$$c(\mathbf{A}_q \Rightarrow \text{Class } h) = \frac{\sum_{\mathbf{x}_p \in \text{Class } h} \mu_{\mathbf{A}_q}(\mathbf{x}_p)}{\sum_{p=1}^{m} \mu_{\mathbf{A}_q}(\mathbf{x}_p)}, \quad h = 1, 2, ..., M, \tag{2}$$

where $\mu_{\mathbf{A}_q}(\mathbf{x}_p)$ is the compatibility grade of $\mathbf{A}_q$. The product operator is used to calculate the compatibility grade of each training pattern $\mathbf{x}_p$ with the antecedent part $\mathbf{A}_q$ of the fuzzy rule R_q in (1). The consequent class C_q is specified as the class with the maximum confidence:

$$c(\mathbf{A}_q \Rightarrow \text{Class } C_q) = \max_h \{c(\mathbf{A}_q \Rightarrow \text{Class } h)\} . \tag{3}$$

The rule weight is specified by the difference between the confidence of the consequent class and the sum of the confidences of the other classes as follows:

$$CF_q = c(\mathbf{A}_q \Rightarrow \text{Class } C_q) - \sum_{\substack{h=1 \\ h \neq C_q}}^{M} c(\mathbf{A}_q \Rightarrow \text{Class } h) . \tag{4}$$

In this manner, the consequent class and the rule weight of each fuzzy rule can be easily determined from training patterns (for details, see [13]).

Using the above-mentioned heuristic manner, we can generate a large number of short fuzzy rules as candidate rules in multiobjective fuzzy rule selection. In our former studies [8, 9], we used the SLAVE criterion [14] to prescreen candidate rules. The use of this criterion, however, tends to exclude many specific and accurate rules. In this paper, we use two threshold values: the minimum *support* and the minimum *confidence*. The fuzzy version of support can be written as

$$s(\mathbf{A}_q \Rightarrow \text{Class } h) = \frac{1}{m} \sum_{\mathbf{x}_p \in \text{Class } h} \mu_{\mathbf{A}_q}(\mathbf{x}_p) . \tag{5}$$

We exclude fuzzy rules that do not satisfy these two thresholds. Among short fuzzy rules satisfying these two thresholds, we choose a prespecified number of the best candidate rules for each class with respect to $s(R_q) \cdot c(R_q)$.

2.1 Multiobjective Genetic Fuzzy Rule Selection

In the second stage, we use the NSGA-II algorithm [15] to generate non-dominated ensemble classifiers from candidate rules extracted in the first stage. To represent an ensemble classifier as a string, we use a multi-classifier coding scheme. Each ensemble classifier is represented by an integer string of length N as $S = s_1 s_2 \cdots s_N$ where N is the number of candidate rules and s_j is an integer (s_j denotes the base classifier to which the jth candidate rule is included). For example, a string "2110201033" denotes an ensemble classifier with three base classifiers: Classifier 1 with R_2, R_3, R_7, Classifier 2 with R_1, R_5, and Classifier 3 with R_9, R_{10}. As shown in this example, the jth candidate rule is not used in any base classifier when $s_j = 0$.

Each ensemble classifier S is evaluated by its accuracy and the diversity of its base classifiers. We use the following two objective functions:

$f_1(S)$: Classification rate of the ensemble classifier S on training patterns,
$f_2(S)$: A diversity measure to evaluate the diversity of the base classifiers in S.

Let S_i be the ith base classifier in the ensemble classifier S. When an input pattern $\mathbf{x}_p$ is to be classified by S_i, a single winner rule R_w is chosen from S_i as

$$\mu_{\mathbf{A}_w}(\mathbf{x}_p) \cdot CF_w = \max\{\mu_{\mathbf{A}_q}(\mathbf{x}_p) \cdot CF_q \mid R_q \in S_i\} . \tag{6}$$

When multiple rules with different consequent classes have the same maximum value in (6), the classification of $\mathbf{x}_p$ by S_i is rejected. The classification of $\mathbf{x}_p$ by S_i is also rejected when there are no compatible rules with $\mathbf{x}_p$ in S_i. The final classification

of $\mathbf{x}_p$ by the ensemble S is performed through the majority voting (strictly speaking "plurality voting") based on the classification result by each base classifier S_i. When multiple classes have the same maximum number of votes, the final classification of $\mathbf{x}_p$ is rejected in this paper whereas random tie-break was employed in [8, 9].

2.2 Various Diversity Measures of Base Classifiers

Various diversity measures have been proposed in the literature [3, 4]. We can use such a diversity measure as the second objective $f_2(S)$ in the previous subsection. In this subsection, we explain the entropy-based diversity measure in our former studies [8, 9] and five other diversity measures in [4], which are examined in the next section.

Entropy based-diversity measure (E): This diversity measure was used in [8, 9] where the entropy E of classification results by base classifiers was calculated as

$$E = \frac{1}{m}\sum_{p=1}^{m}\sum_{c=1}^{M}(-P_{pc}\log P_{pc}). \tag{7}$$

In this formulation, m is the number of patterns, M is the number of classes, and P_{pc} is the ratio of base classifiers which classify the pattern $\mathbf{x}_p$ as Class c. Let us assume that the number of base classifiers in an ensemble classifier is $N_{\text{classifier}}$. If three base classifiers classify the first pattern as Class 1, P_{11} is calculated as $3/N_{\text{classifier}}$. Of course, a larger value of the entropy E means a larger diversity of base classifies.

Disagreement measure (dis): The disagreement measure between two base classifiers S_j and S_k is defined as

$$dis_{j,k} = \frac{n(1,-1)+n(-1,1)}{n(1,1)+n(-1,1)+n(1,-1)+n(-1,-1)}, \tag{8}$$

where $n(1, -1)$ is the number of training patterns that are correctly classified by S_j and misclassified by S_k (i.e., "1" and "−1" represent *true* and *false*, respectively). $n(1, -1)$, $n(1, 1)$ and $n(-1, -1)$ are defined in the same manner as $n(1, -1)$. The disagreement measure among more than two base classifiers is defined as

$$dis = \frac{2}{L(L-1)}\sum_{j=1}^{L-1}\sum_{k=j+1}^{L} dis_{j,k}, \tag{9}$$

where L is the number of base classifiers. A larger value of this disagreement measure means a larger diversity of base classifiers.

Double-fault measure (DF): The double-fault measure for two base classifiers S_j and S_k is the ratio of training patterns that are misclassified by both classifiers:

$$DF_{j,k} = \frac{n(-1,-1)}{n(1,1)+n(-1,1)+n(1,-1)+n(-1,-1)}. \tag{10}$$

The double-fault measure for more than two base classifiers is calculated as

$$DF = \frac{2}{mL(L-1)}\sum_{j=1}^{L-1}\sum_{k=j+1}^{L} DF_{j,k}, \tag{11}$$

where m is the number of training patterns, and L is the number of base classifiers. A smaller value of the double-fault measure means a large diversity of base classifiers.

Kohavi-Wolpert variance (*KW*): The Kohavi-Wolpert variance is calculated as

$$KW = \frac{1}{mL^2}\sum_{i=1}^{m} l_i(L-l_i)\,, \tag{12}$$

where l_i is the number of base classifiers that misclassify the training pattern $\mathbf{x}_i$. A larger value of this measure means a larger diversity of base classifiers.

Measurement of interrater agreement (κ): This measure is calculated as

$$\kappa = 1 - \frac{\sum_{i=1}^{m} l_i(L-l_i)}{mL(L-1)P(1-P)}\,, \tag{13}$$

where P is the average classification rate over L base classifiers. That is,

$$P = 1 - \frac{\sum_{i=1}^{m} l_i}{mL}\,. \tag{14}$$

A smaller value of κ means a larger diversity of base classifiers.

Measure of difficulty (*diff*): The measure of difficulty is defined as the variance of the average classification rate V_i of L base classifiers for each training pattern $\mathbf{x}_p$:

$$diff = \mathrm{var}(V_i)\,, \tag{15}$$

where $V_i = (L-l_i)/L$ for each pattern $\mathbf{x}_i$. A smaller value of this difficulty measure means a larger diversity of base classifiers.

3 Computational Experiments

Our computational experiments were performed for five data sets in Table 1. Attribute values were normalized to real numbers in the unit interval [0, 1]. We used the 10-fold cross validation procedure for performance evaluation of ensemble classifiers.

In the candidate rule extraction phase, we specified the minimum confidence and support, and the upper bound on the number of antecedent conditions as 0.6, 0.01, and

Table 1. Five data sets from the UCI Machine Learning Repository

Data set	Attributes	Patterns	Classes
Breast W	9	683*	2
Diabetes	8	768	2
Glass	9	214	6
Iris	4	150	3
Sonar	60	208	2

* Incomplete patterns with missing values are not included.

three (two only for the Sonar data), respectively. From qualified fuzzy rules, we chose the best 300 rules for each class using the product of support and confidence.

In the genetic rule selection phase, we used the following parameters in NSGA-II:

Population size: 200 strings,
Number of classifiers: 5,
Crossover probability: 0.8,
Biased mutation probabilities:

$$p_m(0 \to a) = 1/300M \quad \text{and} \quad p_m(a \to 0) = 0.1 \quad \text{where} \quad a \in \{1, 2, ..., 5\},$$

Stopping condition: 5000 generations.

Experimental results by a single run of NSGA-II for the Breast W data set are summarized in Fig. 2. Figure 2 shows the relation between the classification rate of each non-dominated ensemble classifier and its value of each diversity measure. In Fig. 2, a large number of non-dominated ensemble classifiers were obtained by a single run of NSGA-II. We can observe a clear tradeoff relation between the accuracy of each ensemble classifier for training patterns and the diversity of its base classifiers. In some plots (e.g., Fig. 2 (a)), we can observe the overfitting of ensemble classifiers. This may suggest the existence of an appropriate value for each diversity measure.

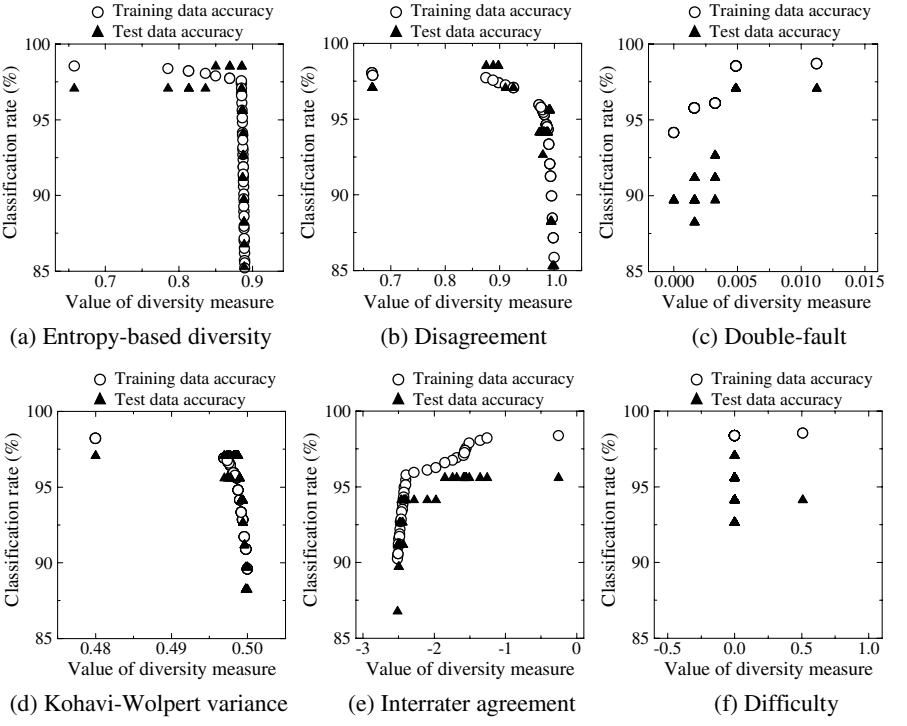

Fig. 2. The accuracy of each ensemble classifier and its diversity (Breast W data)

Table 2 shows the average test data accuracy of the ensemble classifier with the best *training* data accuracy. The highest classification rate on test data for each data set was obtained in each row in Table 2 by a different diversity measure. On the other hand, Table 3 shows the average test data accuracy of the ensemble classifier with the best *test* data accuracy. In almost all runs of NSGA-II, the ensemble classifier with the highest training data accuracy was different from the ensemble classifier with the highest test data accuracy. By the diversity measure based on interrater agreement (κ), the best results on test data were obtained for many data sets.

Table 4 shows the standard deviation of the normalized values of each diversity measure. Before calculating the standard deviation, values of each diversity measure were normalized into real numbers in [0, 1] by its minimum and maximum values over ten runs in the 10-fold cross-validation procedure. When the standard deviation is small, the diversity measure seems to be reliable in the design of ensemble classifiers. In the experimental results in Table 4, the measurement of interrater agreement

Table 2. Average classification rate (%) on test data of the ensemble classifier with the best training data accuracy

	E	*dis*	*DF*	*KW*	κ	*diff*
Breast W	96.04	95.90	**96.48**	94.87	95.61	95.75
Diabetes	**75.13**	74.35	74.62	**75.13**	74.48	74.74
Glass	62.12	59.26	62.12	61.21	**63.48**	60.69
Iris	96.00	94.00	94.67	**96.67**	96.00	95.33
Sonar	74.90	72.55	73.00	73.88	72.07	**75.31**

Table 3. Average classification rate (%) on test data of the ensemble classifier with the best test data accuracy

	E	*dis*	*DF*	*KW*	κ	*diff*
Breast W	97.07	97.21	97.21	96.33	**97.22**	96.48
Diabetes	**77.21**	76.44	76.18	76.70	76.44	76.70
Glass	64.42	63.44	63.51	64.87	**70.00**	63.44
Iris	97.33	96.00	96.00	96.67	**98.67**	96.00
Sonar	80.71	74.95	75.43	76.33	**81.17**	79.64

Table 4. Standard deviation of the normalized values of diversity measure

	E	*dis*	*DF*	*KW*	κ	*diff*
Breast W	0.154	0.352	0.012	0.165	0.119	0.045
Diabetes	0.247	0.100	0.054	0.142	0.157	0.099
Glass	0.277	0.198	0.27	0.257	0.034	0.182
Iris	0.006	0.017	0.264	0.306	0.109	0.000
Sonar	0.09	0.138	0.179	0.268	0.156	0.280
Average	0.155	0.161	0.156	0.228	**0.115**	0.121

(κ) seems to be reliable to specify the diversity of base classifiers. This observation, however, was based on only computational experiments for ensembles with five base classifiers. More computational experiments are needed in future work.

4 Conclusions

In this paper, we examined the use of various diversity measures in our EMO approach to the design of ensemble classifiers. Experimental results showed that the performance of designed ensemble classifiers depended on the choice of a diversity measure. Good results were obtained by the measurement of interrater agreement (κ).

Acknowledgements

This work was partially supported by Grant-in-Aid for Scientific Research on Young Scientists (B): KAKENHI (18700228).

References

1. Breiman, L.: Bagging Predictors. Machine Learning 24, 123–140 (1996)
2. Freund, Y., Schapire, R.E.: A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. Journal of Computer and System Sciences 55, 119–139 (1997)
3. Kuncheva, L.I., Whitaker, C.J.: Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. Machine Learning 51, 181–207 (2003)
4. Tang, E.K., Suganthan, P.N., Yao, X.: An Analysis of Diversity Measures. Machine Learning 65, 247–271 (2006)
5. Abbass, H.A.: Pareto Neuro-evolution: Constructing Ensemble of Neural Networks using Multi-objective Optimization. In: Proc. of 2003 IEEE Congress on Evolutionary Computation, pp. 2074–2080 (2003)
6. Jin, Y., Okabe, T., Sendhoff, B.: Evolutionary Multi-objective Optimization Approach to Constructing Neural Network Ensembles for Regression. In: Coello, C.A.C., Lamont, G.B. (eds.) Applications of Multi-Objective Evolutionary Algorithms, pp. 653–673. World Scientific, Singapore (2004)
7. Chandra, A., Yao, X.: DIVACE: Diverse and Accurate Ensemble Learning Algorithm. In: Yang, Z.R., Yin, H., Everson, R.M. (eds.) IDEAL 2004. LNCS, vol. 3177, pp. 619–625. Springer, Heidelberg (2004)
8. Nojima, Y., Ishibuchi, H.: Designing Fuzzy Ensemble Classifiers by Evolutionary Multiobjective Optimization with an Entropy-based Diversity Criterion. In: Proc. of 6th International Conference on Hybrid Intelligent Systems and 4th Conference on Neuro-Computing and Evolving Intelligence CD-ROM (4 pages) (2006)
9. Nojima, Y., Ishibuchi, H.: Genetic Rule Selection with a Multi-Classifier Coding Scheme for Ensemble Classifier Design. International Journal of Hybrid Intelligent Systems 4(3), 157–169 (2007)
10. 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, 260–270 (1995)

11. 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, 135–150 (1997)
12. Ishibuchi, H., Nakashima, T., Murata, T.: Three-objective Genetics-based Machine Learning for Linguistic Rule Extraction. Information Sciences 136, 109–133 (2001)
13. Ishibuchi, H., Nakashima, T., Nii, M.: Classification and Modeling with Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining. Springer, Berlin (2004)
14. Ishibuchi, H., Yamamoto, T.: Comparison of Heuristic Criteria for Fuzzy Rule Selection in Classification Problems. Fuzzy Optimization and Decision Making 3, 119–139 (2004)
15. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE Trans. on Evolutionary Computation 6, 182–197 (2002)