On the Combination of Accuracy and Diversity Measures for Genetic Selection of Bagging Fuzzy Rule-based Multiclassification Systems

Krzysztof Trawiński, Arnaud Quirin, Oscar Cordón

European Centre for Soft Computing. Mieres, Spain
{krzysztof.trawinski, arnaud.quirin, oscar.cordon}@softcomputing.es

Abstract

A preliminary study combining two choices of a diversity measure with an accuracy measure in two bicriteria fitness functions to genetically select fuzzy rule-based multiclassification systems is conducted in this paper. The fuzzy rule-based classification system ensembles are generated by means of bagging and mutual information-based feature selection. Several experiments were developed using four popular UCI datasets with different dimensionality in order to analyze the accuracy-complexity trade-off obtained by a genetic algorithm considering the two fitness functions. Comparison are made with the initial fuzzy ensemble and a single fuzzy classifier.

1. Introduction

Multiclassification systems (MCSs[1]) are very promising tools to obtain better performance than a single classifier dealing with complex classification problems, especially when the number of dimensions or the size of the data are really large [13]. The most common base classifiers are decision trees [11], neural networks [18], and more recently fuzzy classifiers [3, 20].

In previous studies [5, 6], we described a methodology in which classical MCS design approaches such as bagging [2], random subspace [11], and mutual information-based feature selection [1] are used to generate fuzzy rule-based multiclassification systems (FRBMCSs). The approach is based on a basic heuristic fuzzy classification rule generation method [12] and a classifier selection technique based on a genetic algorithm (GA) driven by a multicriteria fitness function.

We concluded that a feature and an instance selection procedure combined with a simple grid partitioning fuzzy rule-based classification system (FRBCS) to form FRBMCS is a good approach to overcome the curse of dimensionality in large datasets. Nevertheless, once a set of classifiers has been trained, we still need to deal with the high number of rules and the correlations between individual classifiers. This is why a selection of the classifiers is so crucial. As said, we already proposed a multicriteria GA guided by several fitness functions, based on the likelihood [5], the training error [6], or the Out-Of-Bag error [4]. This methodology, quite novel in this topic, lead us to the generation of different compact sets of individual classifiers, while still preserving its accuracy, in a single GA run. However, the experimentations carried suggested the choice of the fitness function is very dependent of the problem being solved. For instance, when using the training error, the accuracy of two FRBMCSs can be similar or even perfect, making difficult for the GA to discriminate between them in order to improve the generalization ability. On the contrary, only using the likelihood seems to give bad results on many datasets. This suggested us to combine different kinds of error-based criteria to overcome this issue, producing better results than any single criterion in isolation [24].

The aim of the current contribution is to take a step ahead in the latter approach by analyzing the influence of diversity measures, which aim to maximize the instability of the individual classifiers composing the MCS to obtain performance improvement. To do so, we propose to combine two diversity measures with the training error to define two different bicriteria fitness functions. As in our previous publication [24] we combine them using the two most simple ways: weighted average and lexicographic order (i.e. considering the optimization of a single criterion, and using the second in case of a tie). We aim to check if two fitness functions using diversity measures (DIVs) [14] will perform better in terms of accuracy than a fitness function using the training error in isolation.

A preliminary study will be conducted on small and medium size datasets from the UCI machine learning repository to test the two different fitness functions, each one using a different DIV, in comparison to a single classifier, the

[1] In the following we will use MCS and ensemble as synonyms.

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.123

original FRBMCS, and the GA-selected FRBMCSs using a training error-based fitness function. Several parameter settings for the global approach (e.g. different granularity levels as well as different feature selection methods) will be tested and compared regarding the accuracy and the size of the rule base obtained by a single classifier and the original FRBCS ensemble.

This paper is set up as follows. In the next section, some of the existing GA-based methods to select MCSs are reviewed as well as a brief overview on the use of diversity measures is shown. Sec. 3 recalls our approach for designing FRBMCSs considering bagging and feature selection. Sec. 4 describes the proposed multicriteria GA for component classifier selection with the two new fitness functions based on the use of the two selected DIVs. The experiments developed and their analysis are shown in Sec. 5. Finally, Sec. 6 collects some concluding remarks and future research lines.

2 Background and Related work

2.1 Genetic selection of MCSs

In general, the selection of a subset of classifiers is done using the *overproduce-and-choose strategy* (OCS) [19], in which a large set of classifiers is produced and then selected to extract the best performing subset. GAs are a popular technique within this strategy. In the literature, performance, complexity and DIV measures are usually considered as search criteria. Complexity measures are employed to increase the interpretability of the system whereas DIVs are used to avoid overfitting.

Among the different genetic OCS proposals, we can remark the following ones. In [17], a hierarchical multi-objective GA (MOGA) algorithm, performing feature selection at the first level and classifier selection at the second level, is presented which outperforms classical methods for two handwritten recognition problems. The MOGA allows both performance and diversity to be considered for MCS selection. In [10], a GA is used to select from seven diversity heuristics for k-means cluster-based ensembles and the ensemble size is also encoded in the genome. In the study of Martínez-Muñoz et al. [15], a GA is compared to five other techniques for ensemble selection. Even if the performance of the GA was the worst obtained, they showed that while selecting a small subset of classifiers, the generalization error was significantly decreased. In [9], the authors developed a multidimensional GA to optimize two weight-based models, in which the weights are assigned to each classifier or to each class. They applied their system to six different classifiers (only linear and quadratic classifiers are explored), but on only two small datasets and without comparing to the results obtained on a single classifier. Finally, our own previous studies [5, 6] also consider a multicriteria GA for the ensemble selection in an OCS fashion, with performance (training error) and complexity as criteria to guide the GA. The performance obtained with the initial MCS is outperformed by the ensemble selected by the GA, while the system is simplified. In our current contribution, we will confirm this conclusion by the study of two improved fitness functions mixing the two most used criteria: the accuracy and the complexity of the classifiers. The fitness function will directly incorporate either one accuracy criteria (the training error) or one accuracy criteria combined with a DIV, while the MCS complexity will be implicitly optimized by the considered coding scheme (see Sec. 4).

2.2 Diversity measures

In general, it seems that obtaining a high diversity between classifiers is the aim to be reached, when aiming to achieve performance improvement of MCSs. In the last few years, a group of researchers devoted their attention to the DIVs [14, 21, 22, 23, 25], as they could improve the instability of classifiers. Several DIVs were proposed, however all of them demonstrated similar characteristics.

In Kuncheva et al. [14] ten different DIVs were proposed to investigate their influence on the ensemble accuracy when being considered as the only optimization criterion. The Q-statistic was the most interesting one, as it showed a correlation between the accuracy and the diversity.

In Ruta et al. [21], classifiers were generated using a single measure, either the diversity, including sixteen different DIVs, or the ensemble error. The best results were obtained with the error and DIVs correlated with the error. The experiment indicated that, out of the whole DIVs selected, the correlation coefficient and the Q-statistic provided the worst results.

Although both authors substituted accuracy by diversity, Tsymbal et al. [25] combined these two measures for feature selection in MCSs and conducted experiments over five different DIVs.

In Dos Santos et al. [22], an experimentation concerning twelve different DIVs used with a single and a multi-objective GA were conducted. Moreover, in [23], four selected DIVs were used to justify a dynamic OCS strategy for the selection of clasifier ensembles. The two best measures introduced in [22, 23] were the double fault and the difficulty.

All the previous authors agreed that DIVs are not useful to substitute the ensemble error, as the correlations depend on the dataset. However, combining a DIV with an error measure is still an open issue, since the use of the latter in isolation seems to be better in most of the cases. This idea led us to include two DIVs into the bicriteria fitness function of our genetic MCS selection method, in combination with the selected ensemble training error.

3 Bagging and feature selection-based FRBMCSs

In this section we will both detail how the individual classifiers and the FRBMCSs are designed. A normalized dataset is split into two parts, a training set and a test set. The training set is submitted to an instance selection and a feature selection procedure in order to provide individual training sets (the so-called *bags*) to train simple FRBCSs (through the method described in Sec. 3.1). The instance selection and the feature selection procedures are described in Sec. 3.2. After performing the training stage on all the bags, we got an initial FRBMCS, which is validated using the training and the test errors as well as a measure of complexity based on the total number of rules in the FRBCSs. This ensemble is selected using a multicriteria GA (described in Sec. 4) guided by accuracy- and diversity-based fitness functions. The final FRBMCS is validated using different accuracy (training error, test error) and complexity (number of classifiers, total number of rules) measures.

3.1 Individual FRBCS design method

The FRBCSs considered in the ensemble will be based on fuzzy rules R_j with a class C_j and a certainty degree CF_j in the consequent: If x_1 is A_{j1} and $\ldots$ and x_n is A_{jn} then Class C_j with CF_j, $j = 1, 2, \ldots, N$. They will take their decisions by means of the single-winner method [12].

To derive the fuzzy knowledge bases, one of the heuristic methods proposed by Ishibuchi et al. in [12] is considered. The consequent class C_j and certainty degree CF_j are statistically computed from all the examples located in a specific fuzzy subspace $D(A_j)$. C_j is computed as the class h with maximum confidence according to the rule compatible training examples $D(A_j) = \{x_1, \ldots, x_m\}$: $c(A_j \Rightarrow Class\ h) = |D(A_j) \bigcap D(Class\ h)|/|D(A_j)|$. CF_j is obtained as the difference between the confidence of the consequent class and the sum of the confidences of the remainder (called CF_j^{IV} in [12]).

3.2 FRBMCS design approaches

The generation of the FRBMCSs is performed by means of a bagging approach combined with a feature selection method [6]. Three different feature selection methods, random subspace [11] and two variants of Battiti's MIFS [1] (greedy and GRASP [8]), are considered.

Random subspace is a method in which we select randomly a set of features from the original dataset. The Battiti's MIFS method is based on a forward greedy search using the Mutual Information measure, with regard to the class. This method selects the set S of the most informative features about the output class which cannot be predicted with the already selected features. It uses a coefficient, β, to set up the penalization on the information brought by the already selected features.

The MIFS-GRASP variant is an approach where the set is generated by iteratively adding features randomly chosen from a Restricted Candidate List composed of the best τ percent decisions according to the Battiti's quality measure. Parameter τ is used to control the amount of randomness injected in the MIFS selection. With $\tau = 0.5$, we get an average amount of randomness, while still preserving the quality-based ordering of the features.

For the bagging approach, the bags are generated with the same size as the original training set, as commonly done. In every case, all the classifiers will consider the same fixed number of features.

Finally, no weights will be considered to combine the outputs of the component classifiers to take the final FRBMCS decision, but a pure voting approach will be applied: the ensemble class prediction will directly be the most voted class in the component classifiers output set. The lowest-order class would be taken in the case of a tie.

4 A multicriteria GA-based MCS selection method

In this section we will report the foundations of the multicriteria genetic selection process. Then, we will introduce the used evaluation criteria and the two new bicriteria fitness functions.

4.1 Multicriteria genetic optimization

The GA searches for an optimal sequence of the classifiers, in the way that the most significant classifiers have the lowest indexes, while those redundant members, which can be safely excluded, are in the last positions. The coding scheme is thus based on an order-based representation, a permutation $\Pi = \{j_1, j_2, \ldots, j_l\}$ of the l originally generated individual classifiers. In this way, each chromosome encodes l different solutions to the problem, based on considering a "basic" MCS comprised by a single classifier, that one stored in the first gene, then another one composed of two classifiers, those in the first and the second genes, and so on.

So, the computation of the evaluation criteria for the whole ensemble is obtained in a *cumulative* way, defined as a vector containing the measured values of the first classifier; the subset formed by the first and the second; and so on. The fitness function is thus using the values of a multicriteria vector, being composed of an array of l values, $L^i = L'_{\{j_1,j_2,\ldots,j_i\}}$, corresponding to the cumulative measure-value of the l mentioned MCS designs. The two different vectors corresponding to two different chromosomes are compared by the highest values of one of the selected criteria (see Sec. 4.2).

At the end of the GA run, the best chromosome is that member in the population overcoming the others using the considered criterion. Then, the final design encoded in this chromosome is the MCS comprising the classifiers from the first to the one having the the best cumulative measured value. Nevertheless, any other design not having the optimal accuracy but, for example, showing a lowest complexity can also be directly extracted. In this way, an implicit use of a complexity criterion is also made.

To increase its convergence rate, the GA works following a steady-state approach. The initial population is composed of randomly generated permutations. In each generation, a tournament selection of size 3 is performed, and the two winners are crossed over to obtain a single offspring that directly substitutes the loser. In this study, we have considered OX crossover and the usual exchange mutation [16].

4.2 The three used evaluation criteria

An exhaustive study using DIVs was conducted in [22, 23]. Two measures can be highlighted from it: the difficulty (θ) and the double fault (δ). In this contribution we have chosen these two DIVs to perform our preliminary study. Apart from θ and δ, we use the Training Error (TE) as the evaluation criteria for the definition of the fitness functions.

The TE is computed as follows. Let $h_1(\mathbf{x}), \ldots, h_l(\mathbf{x})$ be the outputs of the component classifiers of the selected ensemble E for an input value $\mathbf{x} = (x_1, \ldots, x_n)$. For a given sample $\{(\mathbf{x}^k, C^k)\}_{k\in\{1\ldots m\}}$, the TE of that MCS is:

$$TE = \frac{1}{m} \cdot \#\{k \mid C^k \neq \underset{j\in\{1\ldots|E|\}}{\arg\max}\ h_j(\mathbf{x}^k)\} \qquad (1)$$

with $|E|$ being the number of classifiers in the selected ensemble.

Fitness evaluation using TE alone was already studied in one of our previous publications [6]. We will call it *Training Error-based Fitness Function* (TEFF).

The difficulty measure θ is computed as follows. Let $X = \{i/|E|\}_{i\in\{0,\ldots,|E|\}}$ and $X_k \in X$ be the proportions of classifiers classifying correctly the instance x_k. Then, θ is equal to $Var(\{X_1, \ldots, X_k, \ldots, X_m\})$.

The pairwise measure δ for two classifiers h_i and h_j is computed as follows:

$$\delta_{i,j} = \frac{N_{ij}^{00}}{\#samples} \qquad (2)$$

with N_{ij}^{00} being the number of examples missclassified by both h_i and h_j. The global value of the measure for the whole selected ensemble is computed as follows:

$$\delta_{avg} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \delta_{i,j} \qquad (3)$$

with L being the number of component classifiers in the ensemble.

4.3 The two bicriteria fitness functions

As said, we propose two approaches for the fitness function combining one of the selected DIVs (θ, δ_{avg}) with the TE measure, using the *Lexicographical Order-based Fitness Function* (LOFF) or the *Weighted Combination Fitness Function* (WCFF).

Notice that, working in this way, we introduce a second multicriteria optimization level in our algorithm. On the one hand, a multicriteria optimization is made by means of the considered coding scheme and the cumulative evaluation of the possible MCS designs (see Sec. 4.1). On the other hand, a higher level is added when evaluating the latter possible designs by means of a bicriteria fitness function.

In the first one, the LOFF, we use the lexicographical order to deal with the multicriteria optimization. When comparing two chromosomes, one is better than the other if it takes a better (lower) minimum value of the TE. In case of a tie, the DIV measure is considered. The ordering scheme gives priority to TE, as it provided better results in our previous studies, while taking the DIV only as a last resort in the case of the frequent ties encountered by the system.

In the second approach, the WCFF, we propose objective function scalarization by a weighted combination of both measures:

$$WC = factor_0 * \alpha * TE + (1 - \alpha) * DIV \qquad (4)$$

where α is a weight in [0,1] and $factor_0 = DIV_0/TE_0$ is a first evaluation-based normalization using DIV_0 and TE_0, the DIV and the TE values obtained by the evaluation of the initial FRBMCS. The fitness function has to be minimized.

5 Experiments and analysis of results

To evaluate the performance of the generated FRBMCSs, we have selected four datasets from the UCI machine learning repository (see Table 1). In order to compare the accuracy of the considered classifiers, we used Dietterich's 5×2-fold cross-validation (5×2-cv), which is considered to be superior to paired k-fold cross validation in classification problems [7].

Three different granularities, 3, 5 and 7, are tested for the single FRBCS derivation method, for feature sets of size 5

Table 1. Data sets considered

Data set	#attr.	#examples	#classes
Pima	8	768	2
Glass	9	214	7
Vehicle	18	846	4
Sonar	60	208	2

selected by means of three approaches: the greedy Battiti's MIFS filter feature selection method, the Battiti's method with GRASP (with τ equal to 0.5, see Sec. 3.2), and random subspace. Battiti's method has been run by considering a discretization of the real-valued attribute domains in ten parts and setting the β coefficient to 0.1.

The FRBMCSs generated are initially comprised by 50 classifiers. The GA for the component classifier selection works with a population of 50 individuals and runs during 50 generations. The mutation probability considered is 0.05. The weights of WCFF were set to 0.8 for TE and 0.2 for DIV as our aim was to allow a small influence of the DIV in the cases in which the TE gives similar values. The other tested values for the weights did not improve the results significantly.

The statistics (5×2-cv error, number of classifiers, number of rules, and run time required for each run, expressed in seconds) for the genetically selected FRBCS ensembles using LOFF with θ and δ, WCFF with θ and δ, and TEFF are collected in Tables 2 and 3, Tables 4 and 5, and Table 6 respectively. The results of the single FRBCSs are presented in Table 7 while those of the original FRBMCSs are included in Table 8. There are three subtables for each of the feature selection method considered. The best results for a given feature selection methods are shown in bold and the best values overall are outlined.

All the experiments have been run in a linux cluster at the University of Granada on Intel quadri-core Pentium 2.4 GHz nodes with 2 GBytes of memory.

Table 2. Results for the FRBCS ensembles selected by the GA using the LOFF with θ

		Bagging + Greedy				Bagging + GRASP $\tau = 0.50$				Bagging + Random Subspace			
		Pima	Glass	Vehicle	Sonar	Pima	Glass	Vehicle	Sonar	Pima	Glass	Vehicle	Sonar
	5×2-cv	0.252	$\boxed{\mathbf{0.362}}$	0.465	**0.235**	0.251	$\boxed{\mathbf{0.362}}$	0.446	0.232	**0.256**	**0.370**	0.429	$\boxed{\mathbf{0.210}}$
3 labels	#classifiers	4.1	8.2	10.1	9.4	4.0	9.1	11.3	14.0	4.2	11.9	12.6	21.9
5 attr.	#rules	715.5	1018.4	1426.9	1430.1	695.4	1156.7	1705.4	2256.9	674.5	1346.9	2080.4	3645.8
	avg. #rules	175.1	123.5	139.9	151.4	172.8	124.8	153.7	160.3	160.5	114.2	167.0	166.9
	time	1538.42	392.79	2066.12	384.18	1540.23	397.09	1248.16	394.00	1608.29	418.50	1272.52	408.69
	5×2-cv	$\boxed{\mathbf{0.237}}$	0.375	0.395	0.246	**0.244**	0.368	0.403	**0.230**	0.260	0.389	0.380	0.220
5 labels	#classifiers	12.2	11.0	12.8	17.1	11.5	13.8	13.9	26.5	12.6	14.7	14.5	20.4
5 attr.	#rules	7076.6	2959.0	6129.1	9224.6	7074.6	3690.2	8023.7	15862.8	6929.3	3524.1	10290.2	12668.2
	avg. #rules	594.5	269.8	484.1	550.2	608.6	271.8	591.2	601.1	549.9	241.8	725.8	623.5
	time	1540.64	381.62	1992.08	364.14	1548.99	386.68	2039.25	362.66	1611.73	418.61	2087.41	369.85
	5×2-cv	0.251	0.390	**0.369**	0.258	0.252	0.390	$\boxed{\mathbf{0.326}}$	0.240	0.276	0.392	**0.335**	0.263
7 labels	#classifiers	13.5	9.3	14.2	20.3	15.3	11.5	13.5	17.4	16.3	13.8	20.4	13.2
5 attr.	#rules	17537.5	3608.8	15875.1	20283.7	20183.3	4453.6	18818.1	18849.4	20124.0	5276.9	31914.5	14795.2
	avg. #rules	1306.5	390.5	1126.7	998.4	1326.9	399.5	1276.0	1085.4	1225.5	386.5	1585.7	1124.0
	time	1523.94	394.80	1967.13	348.11	1525.76	393.38	1791.23	348.18	1561.90	411.57	2000.05	351.93

Table 3. Results for the FRBCS ensembles selected by the GA using the LOFF with δ

		Bagging + Greedy				Bagging + GRASP $\tau = 0.50$				Bagging + Random Subspace			
		Pima	Glass	Vehicle	Sonar	Pima	Glass	Vehicle	Sonar	Pima	Glass	Vehicle	Sonar
	5×2-cv	0.252	**0.364**	0.466	**0.235**	0.251	$\boxed{\mathbf{0.362}}$	0.443	0.232	**0.256**	**0.376**	0.429	$\boxed{\mathbf{0.210}}$
3 labels	#classifiers	4.1	9.2	10.1	9.4	4.0	9.1	11.8	14.0	4.2	12.2	12.8	21.9
5 attr.	#rules	715.5	1138.5	1414.7	1430.1	695.4	1156.7	1799.0	2256.9	674.5	1356.2	2123.3	3645.8
	avg. #rules	175.1	123.8	139.5	151.4	172.8	124.8	154.1	160.3	160.5	112.7	167.0	166.9
	time	545.14	192.71	677.16	170.28	554.46	204.46	672.80	174.03	555.61	209.39	675.80	175.71
	5×2-cv	$\boxed{\mathbf{0.237}}$	0.375	0.394	0.246	**0.244**	0.368	0.403	**0.230**	0.260	0.389	0.380	0.220
5 labels	#classifiers	12.2	11.0	12.8	17.1	11.5	13.8	13.9	26.5	12.6	14.7	14.5	20.4
5 attr.	#rules	7076.6	2959.0	6129.1	9224.6	7074.6	3690.2	8023.7	15862.8	6929.3	3524.1	10290.2	12668.2
	avg. #rules	594.5	269.8	484.1	550.2	608.7	271.8	591.2	601.1	549.9	241.8	725.8	623.5
	time	538.90	197.21	648.81	163.85	546.70	197.72	649.61	165.04	560.75	202.67	646.24	164.98
	5×2-cv	0.251	0.390	**0.369**	0.258	0.252	0.390	**0.362**	0.240	0.276	0.391	$\boxed{\mathbf{0.334}}$	0.263
7 labels	#classifiers	13.5	9.3	14.2	20.3	15.3	11.5	15.0	17.4	16.3	13.8	20.4	13.2
5 attr.	#rules	17537.5	3608.8	15875.1	20283.7	20183.3	4453.6	20183.4	18849.4	20124.0	5276.9	31914.5	14795.2
	avg. #rules	1306.5	390.5	1126.7	998.4	1326.9	399.5	1367.0	1085.4	1225.5	386.5	1585.7	1124.0
	time	540.56	184.89	636.21	161.39	538.19	185.29	636.57	159.79	547.76	186.37	631.56	163.52

Table 4. Results for the FRBCS ensembles selected by the GA using the WCFF with θ

		Bagging + Greedy				Bagging + GRASP $\tau = 0.50$				Bagging + Random Subspace			
		Pima	Glass	Vehicle	Sonar	Pima	Glass	Vehicle	Sonar	Pima	Glass	Vehicle	Sonar
	5×2-cv	0.252	$\boxed{\mathbf{0.369}}$	0.510	**0.247**	0.248	0.404	0.488	0.250	0.263	0.420	0.455	0.242
3 labels	#classifiers	5.0	14.7	17.5	15.8	1.9	13.8	36.2	20.4	3.2	19.9	43.2	36.0
5 attr.	#rules	848.8	1872.4	2472.8	2394.0	322.0	1795.2	5359.5	3296.4	515.9	2242.9	6916.5	5799.7
	avg. #rules	168.0	120.9	139.7	152.4	168.8	126.2	148.1	164.9	165.0	115.8	160.1	162.1
	time	579.29	194.44	1722.10	172.57	578.00	204.75	1753.04	171.70	581.36	208.72	1777.94	174.94
	5×2-cv	$\boxed{\mathbf{0.236}}$	0.383	0.400	0.249	**0.242**	**0.385**	0.396	$\boxed{\mathbf{0.231}}$	**0.258**	0.405	0.387	**0.238**
5 labels	#classifiers	26.0	20.1	17.5	37.2	15.9	23.2	23.8	33.7	23.5	24.7	30.7	29.0
5 attr.	#rules	15122.4	5188.9	7464.0	19730.2	9194.7	6086.1	12194.1	20080.3	12833.5	6012.9	19284.6	18306.0
	avg. #rules	590.7	266.2	471.7	533.8	583.9	268.0	534.5	597.1	549.9	246.9	669.6	628.3
	time	566.33	197.55	1630.51	164.36	568.78	196.68	1680.45	164.95	587.83	203.96	1731.74	163.93
	5×2-cv	0.249	0.398	**0.374**	0.269	0.254	0.422	**0.357**	0.257	0.267	**0.407**	$\boxed{\mathbf{0.331}}$	0.270
7 labels	#classifiers	27.1	10.0	22.4	30.5	23.8	14.5	17.4	22.7	22.1	26.4	33.6	14.3
5 attr.	#rules	35127.6	3723.4	23698.3	30609.4	31122.7	5798.9	20570.0	24987.1	26650.0	9651.0	46174.9	16376.5
	avg. #rules	1304.2	384.2	1047.3	999.6	1331.3	421.5	1222.0	1092.8	1217.8	370.9	1418.2	1138.6
	time	562.16	185.42	1607.06	161.42	567.65	186.55	1637.21	160.15	574.78	186.83	1644.03	161.52

Table 5. Results for the FRBCS ensembles selected by the GA using the WCFF with δ

		Bagging + Greedy				Bagging + GRASP $\tau = 0.50$				Bagging + Random Subspace			
		Pima	Glass	Vehicle	Sonar	Pima	Glass	Vehicle	Sonar	Pima	Glass	Vehicle	Sonar
	5×2-cv	0.254	0.432	0.504	**0.237**	0.260	0.417	0.481	0.249	0.268	0.411	0.450	0.233
3 labels	#classifiers	5.1	12.9	26.1	19.7	8.3	15.7	22.1	38.4	6.9	18.7	40.6	37.7
5 attr.	#rules	872.1	1581.4	3464.5	2942.4	1436.0	2005.8	3293.3	6102.2	1079.9	2132.2	6475.3	6121.2
	avg. #rules	174.7	123.5	136.9	148.1	176.4	124.1	150.0	158.7	157.2	117.5	159.8	162.6
	time	1214.17	192.91	1722.12	172.12	1218.58	206.23	1757.07	171.14	1287.95	208.39	1777.99	174.37
	5×2-cv	$\boxed{\mathbf{0.234}}$	**0.378**	0.399	**0.237**	**0.237**	**0.383**	0.395	$\boxed{\mathbf{0.217}}$	**0.258**	$\boxed{\mathbf{0.361}}$	0.382	**0.220**
5 labels	#classifiers	27.5	28.2	33.1	41.4	29.8	28.8	22.2	40.0	24.1	25.6	39.7	28.8
5 attr.	#rules	16377.8	7230.0	14789.7	22250.9	17760.7	7715.0	11115.7	24051.3	13411.9	6236.1	24477.2	18050.8
	avg. #rules	591.4	258.4	448.7	535.4	599.5	277.4	517.3	601.8	555.9	220.2	616.3	24.5
	time	1219.26	197.94	1630.61	164.87	1228.92	197.85	1680.90	165.92	1297.38	168.55	1739.54	166.50
	5×2-cv	0.251	0.423	**0.376**	0.255	0.251	0.404	**0.349**	0.240	0.263	0.409	$\boxed{\mathbf{0.331}}$	0.263
7 labels	#classifiers	33.1	30.4	27.0	25.6	26.5	24.0	24.8	20.2	26.4	33.8	47.0	13.2
5 attr.	#rules	42435.9	11567.8	26328.8	25278.1	34799.5	9025.1	27967.1	21950.1	31898.1	12195.5	63657.6	14795.2
	avg. #rules	1291.9	372.8	991.2	991.5	1322.7	379.2	1151.5	1087.0	1206.1	366.0	1353.9	1124.0
	time	1202.92	185.15	1609.29	161.42	1208.79	185.00	1639.83	161.78	1245.55	185.50	1646.18	161.23

Table 6. Results for the FRBCS ensembles selected by the GA using the TEFF

		Bagging+Greedy				Bagging+GRASP $\tau = 0.50$				Bagging + Random Subspace			
		Pima	Glass	Vehicle	Sonar	Pima	Glass	Vehicle	Sonar	Pima	Glass	Vehicle	Sonar
	5×2-cv	0.257	$\boxed{\mathbf{0.360}}$	0.461	**0.235**	0.254	0.372	0.449	**0.237**	**0.256**	**0.381**	0.428	$\boxed{\mathbf{0.216}}$
3 labels	#classifiers	4.1	7.3	10.3	12.3	14.4	10.2	12.9	13.9	4.2	13.7	13.4	20.1
5 attr.	#rules	696.5	904.3	1431.0	1842.1	763.0	1317.9	1991.6	2252.6	703.4	1546.0	2239.5	3376.7
	avg. #rules	171.5	125.4	138.3	148.3	174.3	126.0	155.9	161.7	168.1	113.1	168.9	168.3
	time	94.06	26.35	103.26	25.32	93.37	26.49	102.09	25.18	92.77	26.39	103.24	25.08
	5×2-cv	**0.242**	0.383	0.392	0.247	$\boxed{\mathbf{0.239}}$	**0.363**	0.399	0.252	0.263	0.392	0.378	0.249
5 labels	#classifiers	11.5	15.9	15.5	10.4	10.9	14.7	12.0	7.8	11.9	13.7	13.0	9.4
5 attr.	#rules	6744.9	4233.1	7338.4	5757.7	6497.4	3986.7	7227.3	4893.9	6680.0	3312.2	9455.9	6208.8
	avg. #rules	592.8	268.7	481.9	567.0	593.5	282.0	611.3	630.0	555.8	245.0	734.3	668.8
	time	93.48	26.10	103.48	25.17	92.58	26.16	103.75	24.86	91.47	26.18	104.81	24.83
	5×2-cv	0.258	0.393	**0.374**	0.258	0.256	0.395	**0.356**	0.257	0.265	0.393	$\boxed{\mathbf{0.337}}$	0.267
7 labels	#classifiers	12.7	8.9	14.6	6.3	16.4	10.3	13.2	6.7	17.0	15.5	17.5	6.4
5 attr.	#rules	16614.3	3524.3	16102.3	6427.0	21836.6	4140.6	18296.2	7767.8	21289.5	5980.6	28854.2	7655.2
	avg. #rules	1313.9	404.5	1115.7	1040.9	1346.2	401.9	1386.5	1148.7	11248.4	386.2	1680.2	1203.7
	time	92.87	26.50	102.90	24.85	92.49	26.18	102.93	25.31	92.31	26.08	103.52	25.19

Table 7. Results for the single FRBCSs with feature selection

		Greedy				GRASP $\tau = 0.50$				Random Subspace			
		Pima	Glass	Vehicle	Sonar	Pima	Glass	Vehicle	Sonar	Pima	Glass	Vehicle	Sonar
3 labels	5×2-cv	0.266	0.446	0.549	$\boxed{\mathbf{0.261}}$	0.267	0.447	0.546	0.316	0.265	0.457	0.512	**0.319**
5 attr.	#rules	178.50	135.30	136.40	146.60	179.50	137.00	135.80	169.00	161.80	109.50	154.50	174.50
	time	0.08	0.04	0.12	0.08	0.09	0.04	0.12	0.09	0.07	0.03	0.12	0.08
5 labels	5×2-cv	$\boxed{\mathbf{0.246}}$	**0.376**	0.430	0.287	$\boxed{\mathbf{0.246}}$	$\boxed{\mathbf{0.375}}$	0.425	**0.314**	**0.262**	0.435	0.460	0.329
5 attr.	#rules	682.70	291.00	437.60	615.20	682.70	293.50	418.90	752.70	604.20	259.60	587.80	773.60
	time	0.42	0.25	0.65	0.16	30.39	0.26	0.63	0.17	0.36	0.24	0.67	0.17
7 labels	5×2-cv	0.262	0.414	**0.402**	0.291	0.266	0.423	$\boxed{\mathbf{0.399}}$	0.317	0.276	**0.418**	**0.415**	0.340
5 attr.	#rules	1600	431.20	1021	1218	1599	437.20	907.50	1470	1432	410.90	1266	1536
	time	1.75	1.32	3.27	0.52	1.71	1.34	3.25	0.55	1.66	1.32	3.37	0.63

Table 8. Results for the FRBCS ensembles

		Bagging+Greedy				Bagging+GRASP τ = 0.50				Bagging + Random Subspace			
		Pima	Glass	Vehicle	Sonar	Pima	Glass	Vehicle	Sonar	Pima	Glass	Vehicle	Sonar
3 labels 5 attr.	5×2-cv	0.261	0.463	0.525	0.255	0.262	0.464	0.494	0.246	0.299	0.450	0.453	0.250
	#rules	8578	6208	6843	7282	8609	6289	7362	7951	7936	5671	8008	8174
	avg. #rules	171.55	124.16	136.87	145.65	172.18	125.77	147.24	159.03	158.71	113.42	160.16	163.47
	time	3.43	1.51	4.87	2.52	3.45	1.53	4.91	2.57	3.34	1.49	5.06	2.58
5 labels 5 attr.	5×2-cv	**0.235**	$\boxed{\textbf{0.396}}$	0.400	**0.240**	$\boxed{\textbf{0.234}}$	**0.405**	0.399	$\boxed{\textbf{0.220}}$	**0.260**	0.430	0.378	**0.221**
	#rules	29405	12877	22177	26769	29748	13302	25578	30068	27199	11998	30799	31824
	avg. #rules	588.11	257.54	443.55	535.37	594.95	266.04	511.56	601.36	543.97	239.96	615.97	636.47
	time	17.93	12.11	31.21	6.66	18.05	12.23	32.79	6.96	17.64	11.94	33.91	7.13
7 labels 5 attr.	5×2-cv	0.243	0.430	**0.375**	0.262	0.247	0.425	**0.353**	0.242	0.263	**0.402**	$\boxed{\textbf{0.330}}$	0.241
	#rules	64891	18633	48479	49587	65802	19272	54721	54684	59824	17999	67936	57298
	avg. #rules	11298	372.66	969.58	991.74	1316	385.45	1094	1094	1196	359.98	1359	1146
	time	84.70	67.36	166.51	24.72	85.27	68.27	170.48	25.49	82.12	66.06	174.24	25.57

5.1 Comparison of the diversity measures for the LOFF approach

Comparing the two diversity measures, we can see how they achieve very similar results both in terms of accuracy and complexity:

- while θ is able to outperform δ considering the individual test error 4 times, the latter measure outperforms the former other 4 times, although with less significant differences,
- we should remark the large number of draws (32),
- the best individual improvement was observed on the vehicle dataset: -10% with GRASP and 7 labels,
- the best overall result was obtained on the sonar dataset with Random and 3 labels (draw with δ) and on the vehicle dataset with GRASP and 7 labels,
- concerning the number of selected classifiers, θ achieves a lower value in 5 of the 36 cases, showing the same result in the reminder 31.

Thus, we choose θ as a LOFF representant in the following comparisons (referred simply as the *selected LOFF*).

5.2 Comparison of the diversity measures for the WCFF approach

Comparing the two diversity measures, we can see how θ was outperformed by δ considering the individual test error 25 times (2 draws). However, it achieves lower complexity levels in 27 of the 36 cases. The best individual error improvement was observed on the glass dataset -11% with Random and 5 labels. The best overall result was obtained on the pima dataset with Greedy and 5 labels. Thus, we choose δ as a WCFF representant in the following comparisons (referred simply as the *selected WCFF*).

5.3 Comparison of the three fitness functions

The major observations based on the results are:

- comparing the three fitness functions, we can see how the selected WCFF approach is able to outperform the TEFF and the selected LOFF considering the individual test error 8 times (there were also 3 draws),
- the best overall result was obtained on the sonar dataset with Random and 3 labels and on the vehicle dataset with GRASP and 7 labels,
- however, the FRBMCSs based on the selected LOFF are better than those generated with the TEFF and the selected WCFF in 13 of the 36 cases (apart from 4 draws),
- the best overall result was obtained on the pima dataset with Greedy and 5 labels,
- the TEFF-based FRBMCSs outperform the LOFF and the WCFF considering individual test error for 11 of the 36 times (1 draw). The best overall result was obtained on the glass dataset with Greedy and 3 labels,
- concerning the complexity reduction, TEFF achieves the lowest number of classifiers in 18 of the 36 cases while the selected LOFF does so in the other 16 (apart from 2 draws between them). The selected WCFF always generates the most complex FRBMCSs.

We may conclude that the selected LOFF and WCFF are competitive with TEFF:

- In the direct comparison, the use of the selected LOFF improves the single TEFF performance in 22 out of 36 cases (apart from 2 draws),
- the WCFF improves the single TEFF performance in 16 out of 36 cases,
- which indicates that the joint combination of the TE and a diversity measure actually allows us to improve the performance of the generated FRBMCS in our experimentation.

5.4 Genetically selected FRBMCSs vs. single FRBCS/original FRBMCSs

In all the 36 cases, the generated FRBMCSs improve the performance of the single FRBCS. Besides, although the main goal of the genetic selection is to reduce the complexity of the generated FRBMCS, the accuracy results obtained from that process are also improved in most of the cases, showing the potential of the approach. In only 6 of the 36 cases (apart from 2 draws) the original FRBMCS outperforms the best genetically designed one in terms of accuracy. Comparing the best overall TE values of the genetically selected FRBMCSs with those of the original FRBMCSs, the GA improves the results on three of the considered datasets: vehicle (-1.5% regarding the selected LOFF), glass (-9% regarding the TEFF), and sonar (-4.5% regarding the LOFF with both θ and δ), only giving an equal result in the case of pima.

6 Conclusions and future works

In this study, we extended our previously developed methodology in which a bagging approach together with a feature selection technique are used to train FRBMCSs, which are selected by a multicriteria GA at a later stage. Three fitness functions were tested, the TEFF, the LOFF, and the WCFF, respectively based on a single accuracy criterion and on its combination with a DIV (θ, δ). The selected FRBCS ensembles obtained performed correctly on classification problems with a significant number of features. By using the said techniques, we would like to obtain FRCMCSs dealing with high dimensional data.

One of the next steps we will consider is the design of a generic framework to define the multicriteria fitness function. At least two different information levels will be studied: the chromosome and the objective level. Furthermore, we would like to extend this study on larger data sets (more than 1,000 examples), to study the influence of other parameters (the GA parameters, etc.), and to design more advanced genetic MCS selection techniques (e.g. the use of Pareto-based algorithms). Analysis of other fuzzy rule generation techniques and different diversity criteria in the algorithm are other important points for future research.

References

[1] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 1994.

[2] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[3] J. Canul-Reich, L. Shoemaker, and L. Hall. Ensembles of fuzzy classifiers. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, London, 2007.

[4] O. Cordón and A. Quirin. Comparing two genetic overproduce-and-choose strategies for fuzzy rule-based multiclassification systems generated by bagging and mutual information-based feature selection. *International Journal of Hybrid Intelligent Systems*, 2009. *In press.*

[5] O. Cordón, A. Quirin, and L. Sánchez. A first study on bagging fuzzy rule-based classification systems with multicriteria genetic selection of the component classifiers. In *Third International Workshop on Genetic and Evolving Fuzzy Systems (GEFS)*, pages 11–16, Witten-Bommerholz, 2008.

[6] O. Cordón, A. Quirin, and L. Sánchez. On the use of bagging, mutual information-based feature selection and multicriteria genetic algorithms to design fuzzy rule-based classification ensembles. In *International Conference on Hybrid Intelligent Systems (HIS)*, pages 549–554, Barcelona, 2008.

[7] T. Dietterich. Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

[8] T. Feo and M. Resende. Greedy randomized adaptive search procedures. *Journal of Global Optimization*, 6:109–133, 1995.

[9] B. Gabrys and D. Ruta. Genetic algorithms in classifier fusion. *Applied Soft Computing*, 6(4):337–347, 2006.

[10] S. Hadjitodorov and L. Kuncheva. Selecting diversifying heuristics for cluster ensembles. *Lecture Notes in Computer Science*, 4472:200–209, 2007.

[11] T. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[12] H. Ishibuchi, T. Nakashima, and M. Nii. *Classification and Modeling With Linguistic Information Granules*. Springer, 2005.

[13] L. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, 2004.

[14] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

[15] G. Martínez-Muñoz, D. Hernández-Lobato, and A. Suárez. Selection of decision stumps in bagging ensembles. *Lecture Notes in Computer Science*, 4668:319–328, 2007.

[16] Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer, 1996.

[17] L. Oliveira, M. Morita, R. Sabourin, and F. Bortolozzi. Multi-objective genetic algorithms to create ensemble of classifiers. *Lecture Notes in Computer Science*, 3410:592–606, 2005.

[18] D. Optiz and R. Maclin. Popular ensemble methods: an empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999.

[19] D. Partridge and W. Yates. Engineering multiversion neural-net systems. *Neural Computation*, 8(4):869–893, 1996.

[20] W. Pedrycz and K. Kwak. Boosting of granular models. *Fuzzy Sets and Systems*, 157(22):2934–2953, 2006.

[21] D. Ruta and B. Gabrys. Classifier selection for majority voting. *Information Fusion*, 6(1):63–81, 2005.

[22] E. D. Santos, R. Sabourin, and P. Maupin. Single and multi-objective genetic algorithms for the selection of ensemble of classifiers. In *International Joint Conference on Neural Networks (IJCNN)*, pages 3070–3077, Vancouver, 2006.

[23] E. D. Santos, R. Sabourin, and P. Maupin. A dynamic overproduce-and-choose strategy for the selection of classifier ensembles. *Pattern Recognition*, 41(10):2993–3009, 2008.

[24] K. Trawiński, A. Quirin, and O. Cordón. Bi-criteria genetic selection of bagging fuzzy rule-based multiclassification systems. In *IFSA World Congress-EUSFLAT Conference*, Lisbon, 2009. *In press.*

[25] A. Tsymbal, M. Pechenizkiy, and P. Cunningham. Diversity in search strategies for ensemble feature selection. *Information Fusion*, 6(1):83–98, 2005.