



A selection approach for scalable fuzzy integral combination

P. Bulacio^{a,*}, S. Guillaume^b, E. Tapia^a, L. Magdalena^c

^a Cifasis, Conicet, 27 de febrero 210B, S2000EZP Rosario, Argentina

^b Cemagref, UMR ITAP, BP5095, 34196 Montpellier, France

^c European Centre for Soft Computing, Edf. Científico Tecnológico, Mieres, Spain

ARTICLE INFO

Article history:

Received 1 November 2007

Received in revised form 17 April 2009

Accepted 17 June 2009

Available online 22 June 2009

Keywords:

Multiclassifier scalability

Fuzzy integral

Greedy selection

Cooperative classification

ABSTRACT

We consider the problem of collective decision-making from an arbitrary set of classifiers under the Sugeno fuzzy integral (SFI). We assume that classifiers are given, i.e., they cannot be modified towards their effective combination. Under this baseline, we propose a selection-combination strategy, which separates the whole process into two stages: the classifiers selection, to discover a subset of cooperative classifiers under SFI, and the typical SFI combination of selected classifiers. The proposed selection is based on a greedy algorithm which through a heuristic allows an efficient search.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Multiclassifier systems are aimed at enhancing the performance of any single classifier. Although there are many ways to use more than one classifier, the effectiveness of collective results entails the cooperation among classifiers, i.e., classifiers specifically combined should not propagate individual mistakes to collective results. In particular, cooperation can be easily achieved if classifiers make errors in different samples.

The design of multiclassifier systems usually involves two steps [18]: the generation of classifiers, and their combination. In general, the first step creates a set of diverse classifiers to induce their cooperation towards a later combination [2,11,16]. However, classifiers may be given and just the combination stage can be done [13,14]. In this latter case, the cooperation must be merely exploited without altering the classifier behavior. A typical example of this situation is focused in this paper: the requirement of a single decision-making from a population of classifiers which cannot be adapted/alterd for collective working.

The problem of efficiently combining an arbitrary population of classifiers entails a hard combinatorial problem. Since most of the hypotheses required by untrained combination rules, e.g., the independence assumption [7], cannot be guaranteed, their effectiveness is strongly limited. Alternatively, trained combination rules may be more appropriate owing to their ability to exploit the collective generalization strength of classifier subsets. However, the

induction of such knowledge may be computationally prohibitive, even if a handful of classifiers are considered. In this paper, a greedy approach for the efficient design of trained combination rules over populations of classifiers of arbitrary size is presented. Regarding this objective, the overall combination process is divided into two complementary processes: *selection and combination* (Fig. 1).

At the selection step, the initial set of classifiers is reduced to a tractable subset of cooperative classifiers. Such reduction is accomplished under constraints of efficiency and effectiveness. Regarding efficiency, exhausted searches are avoided by introducing a heuristic search guided by a cooperation ability index. This index evaluates the potential cooperation ability of subsets of classifiers under a given combination rule. Regarding effectiveness, the selected combination rule should be able to deeply characterize the collective behavior of arbitrary subsets of classifiers. In this proposal, the Sugeno integral [15] is considered. The Sugeno integral implements a simple, yet powerful combination mechanism, which takes into account the collective generalization strength of classifiers by means of a fuzzy measure. After the selection, the combination takes place.

The paper is organized as follows. In Section 2, a selection-combination strategy based on Sugeno fuzzy integral is presented. In Section 3, experimental results on benchmark UCI and real data are presented. Finally, in Section 4, conclusions are presented.

2. Selection-SFI combination strategy (sSFI)

The selection of cooperative classifiers should address the following questions: (1) Which are the features governing their effec-

* Corresponding author. Tel.: +54 3414821771; fax: +54 3414821771 52.
E-mail address: bulacio@cifasis-conicet.gov.ar (P. Bulacio).

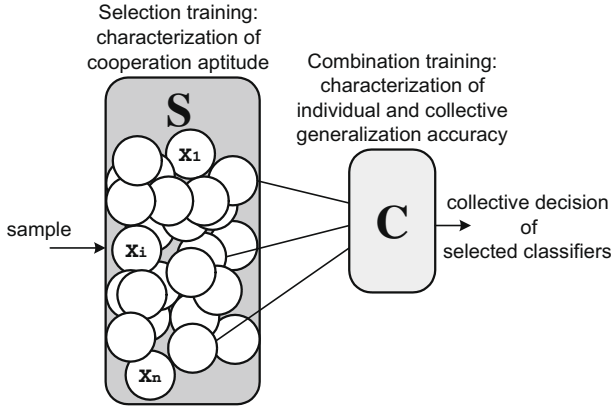


Fig. 1. Selection-combination strategy.

tive work under a posterior and known combination rule? and (2) How to reach a cooperative subset of classifiers in a cost-efficient manner? Considering a Sugeno fuzzy integral combination rule, its behavior must be analyzed to answer the first question. To answer the second one, a heuristic selection that exploits the information of the former step, based on greedy algorithms is suggested.

2.1. Sugeno fuzzy integral combination

The fuzzy integral (FI) is a general trained combination method. Its definition w.r.t. a fuzzy measure [15] provides a good framework to represent imprecise knowledge associated with the behavior of classifier subsets. See [10] for details. We focus on SFI assuming a given population of classifiers, $X = \{X_1, \dots, X_i, \dots, X_n\}$, which associate each input s with $W = \{w_1, \dots, w_j, \dots, w_c\}$ class space. The classification function of the i th classifier is $f_i : s \rightarrow [0, 1]^c$. The f_i components (f_i^1, \dots, f_i^c) can be interpreted as degrees of support of the i th classifier to each class prediction.

Collective FI results are obtained by aggregating levels of decision where classifiers agree, with collective generalization abilities (g) of classifiers that support them. This generalization strength of sets of classifiers is characterized by fuzzy measures, also named fuzzy densities when a single classifier is considered.

A set function $g : 2^X \rightarrow [0, 1]$ is a fuzzy measure if it satisfies the following conditions:

- (1) $g(\emptyset) = 0, g(X) = 1$ (boundary conditions).
- (2) $A \subseteq B \Rightarrow g(A) \leq g(B)$ (monotonicity) $\forall A, B \in 2^X$.

The Sugeno integral [15] of a function $f : X \rightarrow [0, 1]$ w.r.t. g on $(X, 2^X)$ is defined by

$$S_g(f) := \max_{i=1}^n \{ \min(f(X_{(i)}), g(A_{(i)})) \} \quad (1)$$

(i) indicates the permutation of indices, $0 \leq f(X_{(1)}) \leq \dots \leq f(X_{(n)}) \leq 1$, $f(X_{(0)}) := 0$, and $A_{(i)} := \{X_{(1)}, \dots, X_{(i)}\}$. The measure $g(A_{(i)})$ (or $g^{(i)}$ for short) quantifies the generalization ability of the subset $A_{(i)}$. In particular, the SFI for the class w_j is: $S_g(f_j) := \max_{i=1}^n \{ \min(f_{(i)}^j, g_j^{(i)}) \}$.

2.1.1. Behavior of Sugeno integral

When the first step of collective classification design is the construction of classifiers, their collective behavior can be induced. However, when classifiers are externally given, the collective behavior must be carefully analyzed and characterized during the multiclassification procedure.

The collective behavior of classifiers under SFI depends on f and g values, i.e., the relationship among classification decisions and the value of generalization ability determines the final decision. Consequently, two situations can happen:

- (1) The final decision is defined by just one classifier. This happens when there is a classifier with clear decisions (f_i) and strong measures of generalization ability (g^i), i.e., the minimum among a classifier decision and its fuzzy density is larger than the rest of fuzzy densities. The corresponding classifier is named *predominant*.
- (2) The final decision is collectively defined. This situation is presented when there is no classifier that prevails in its decision-ability relation over the others. So, the final result depends on the collective generalization ability of classifiers that agree on different levels of decisions.

Clearly, both situations show that SFI can be successful with a correct predominant or with a correct consensus. Under these conditions, a possible way to exploit the power of SFI is selecting those classifiers that maximize the number of well classified samples in the training dataset, taking into account the f - g relationship.

2.2. Selection process

The proposed selection is based on a heuristic search by a greedy algorithm [1]. The selection is computed based on a single criterion (selection rule), instead of having a recursive analysis over any alternative. The process starts with an empty set to which the most promising candidates are added until no improvement in the overall behavior is obtained (stopping rule).

Being X a set of candidates from which a selection is done, the process involves iterations over the followings steps:

- (1) Classifiers selection: It chooses the *best* candidate for working with the already selected classifiers by applying a selection rule
- (2) Selection end: It determines the contribution of new candidates and the algorithm cut through a stopping rule.

2.2.1. Selection rule

The set of selected classifiers (set O) starts as an empty set that is extended with the best candidate (X_r^*) at each selection step. X_r^* is determined from the analysis of each extended subset, $O_r = \{X_r \cup O\}$ with $r = 1, \dots, n_r$, being n_r the amount of candidates. With this aim, the selection knowledge completes the SFI behavior description: while SFI knowledge handles a full description of the collective behavior (2^X subsets) at the classifier level, the selection knowledge provides a simplified view of collective behavior at the sample level. The selection, through the selection rule, picks the candidate that exploits the cooperation under SFI, i.e., correct consensus or correct predominant. To quantify the potential consensus or predominance among candidates, the following indices are introduced.

- (1) The *coverage index* that evaluates the minimal condition to achieve correct collective results: for each sample, at least one classifier of O_r must be correct.
- (2) The *f - g relationship index* that evaluates the possibility of correct predominant or consensus by the relation among decisions-abilities of O_r classifiers.

In order to achieve correct SFI results, X_r^* should maximize both indices. To facilitate the study of *coverage* and *f - g relationship*, the

matrices of decision pattern F and error pattern E on $Z = \{z_k\}$, with $k = 1, \dots, K$, are analyzed.

$$F, E = \begin{matrix} & F_1, E_1 & \cdots & F_i, E_i & \cdots & F_n, E_n \\ \begin{matrix} z_1 \\ \vdots \\ z_k \\ \vdots \\ z_K \end{matrix} & \left(\begin{matrix} f_1(z_1), 0 & \cdots & f_i(z_1), 1 & \cdots & f_n(z_1), 1 \\ \vdots & & & & \vdots \\ & & f_{k,i}, e_{k,i} & & \\ \vdots & & & & \vdots \\ f_1(z_K), 1 & \cdots & f_i(z_K), 0 & \cdots & f_n(z_K), 0 \end{matrix} \right) \end{matrix}$$

While $e_{k,i} = 1$ means correct classification and $e_{k,i} = 0$ implies error, $f_{k,i}$ is the decision vector of X_i on the sample z_k associated with the class space W . The matrices E and F enclose a complete classifier generalization description and several diversity and accuracy measures [6,9] can be computed from them: their vertical scanning shows the individual generalization strength on Z , and their horizontal scanning shows the collective behavior per sample.

Coverage index of X_r, B_r : It is computed as the average coverage on Z of classifiers of O_r , being the coverage per sample:

$$b_{k,r} = \begin{cases} 0, & \text{if classifiers of } O_r \text{ have a common error on } z_k; \\ 1, & \text{if at least one classifier of } O_r \text{ is correct on } z_k. \end{cases} \quad (2)$$

$b_{k,r}$ values are initialized with the error pattern of the first selected classifier. B_r (with $r = 1, \dots, n_r$) is the fraction of covered samples on Z , i.e., the proportion of “ones” of $b_{k,r}$, with $k = 1, \dots, K$.

f - g relationship value of X_r, FG_r : Being w_j the correct class of the sample z_k , the f - g relationship of X_r values: (i) the coverage strength if X_r is correct in z_k , or (ii) the positive consensus contribution to w_j if X_r is wrong. FG_r is computed as the mean f - g value on the training set Z .

- (i) the coverage strength per sample z_k is the maximal correct decision (class w_j) of O_r members weighted by the generalization ability.

$$s_{k,r} = \max_{q=1}^{Q_r} \left\{ \min \left(f_{q,k}^j, g_j^q \right) \right\} \quad (3)$$

being Q_r the cardinality of O_r .

- (ii) the consensus per sample z_k is the average of correct decision values (class w_j) of O_r members weighted by the generalization abilities.

$$c_{k,r} = \frac{1}{Q_r} \sum_{q=1}^{Q_r} f_{q,k}^j \times g_j^q \quad (4)$$

The selection of X_r^* gives priority to the classifier that maximizes values of decisions-abilities when it is correct, and positive consensus when it is wrong. Based on the above characterizations, a vector of f - g characterization (fg_r of X_r) is built.

$$fg_{r,k} = \begin{cases} s_{k,r} & \text{if } X_r \text{ is correct in } z_k; \\ c_{k,r} & \text{if } X_r \text{ is wrong in } z_k. \end{cases} \quad (5)$$

The f - g relationship value FG_r of each X_r is valued as the mean of the components $fg_{r,k}$.

Selection rule: X_r^* is the candidate that achieves, with the already selected ones, the major coverage and f - g relation on Z . $X_r^* \leftrightarrow \max_{r=1}^{n_r} \{B_r + FG_r\}$

The coverage of O_r is usually stronger than the f - g relation value of X_r in quite good classifiers. Note that the *sum* in the selection rule is the simplest way to discriminate solutions with a similar B . Other solutions, such as hierarchic selections can be considered.

2.2.2. Selection end

A new X_r^* becomes a member of O whenever its selection contributes to the combination. To decide its inclusion, the collective performance P_r^* of the best candidate with the already selected ones is estimated and compared with the performance of already selected classifiers P_O . With this aim, the existence of predominant classifiers is determined. Samples store the f - g relationship, initially, the highest value (associated with w_m class) of minimum f - g relation of X_r , $\min(f_r^m; g_m^r)$. This value is compared with the highest decision (f_r^{m*}) of X_r^* . The following cases can occur in each sample:

- (1) If the $\min(f_r^m; g_m^r) > f_r^{m*}$ then X_r continues predominating.
- (2) If the $\min(f_r^m; g_m^r) < f_r^{m*}$ we can have:
 - If $\min(f_r^{m*}; g_m^{r*}) > f_r^m$ then X_r^* predominates, being g_m^{r*} the f_r^{m*} corresponding fuzzy density.
 - If $\min(f_r^{m*}; g_m^{r*}) < f_r^m$ then there are no predominant. An estimation of the collective performance of O_r^* , such as weighted vote with f, g measures, is required.

If the candidate X_r^* predominates in z_k , the values of the sample characterization are updated by those of X_r^* . In addition, $P_r^*(z_k)$ is directly evaluated by comparing w_{m*} with the real class w_j . Otherwise, collective performance is estimated.

2.2.3. Selection algorithm

The main inputs are the matrices E and F of the given set of classifiers. They are evaluated using *ten*-fold cross-validation on the training set Z .

Process beginning: Given are $E_{K \times n}, F_{K \times n}$.

- (1) Evaluate the individual accuracy of classifiers and select the most accurate as the initial member of O , denoted by X_b .
- (2) Evaluate the **coverage index** $B_r = \frac{1}{K} \sum_{k=1}^K b_{k,r}$ of each X_r with $r = 1, \dots, n_r$ and with $k = 1, \dots, K$.
- (3) Evaluate the **f - g relationship index** FG_r of each X_r according to $fg_{r,k}$ sample values.
- (4) Choose X_r^* applying the **selection rule**: $X_r^* \leftrightarrow \max_{r=1}^{n_r} \{B_r + FG_r\}$.
- (5) Evaluate the **selection end rule**: P_r^* to decide the X_r^* inclusion:
 - IF $(P_r^* < P_O \cdot \alpha)$; $\alpha \in [0, 1]$
 - THEN stop selections.
 - ELSE $O = \{O \cup X_r^*\}$ and GOTO 2.

In the first step, the most accurate classifier (X_b) is included in O . In this way, an initial subset is defined and its greedy augmentation starts. In addition, coverage values $b_{k,r}$ are initialized with its error pattern.

The selection of X_r^* is done according to the potential cooperation among the already selected classifiers and the remaining ones. The cooperation is evaluated using measures of coverage and f - g relationship. These measures are computed using the given error pattern and decision pattern matrices. B_r is an optimist estimation of collective error distribution if the candidate X_r was included in O ; a one entry in $b_{k,r}$ means that at least one classifier of O_r classifies correctly the row sample. Additionally, the f - g relationship characterizes the strength of the candidate contribution depending on its levels of decisions and generalization abilities on Z dataset. A highest level of correct f - g relationship of some classifier of O , as well as the high positive consensus, may give a correct sample classification even if the new candidate is mistaken.

The selection process continues until the collective performance drops. The parameter α prevents the method for possible staking, especially at the beginning where the best classifier could reject further inclusions. We should note that the cooperation is sometimes impossible, e.g., when one classifier is much better than others. In that case the combination is not proper and the use of the best classifier is better.

3. Experiments

We evaluate selection-SFI combination approach on benchmark UCI and real datasets using 10 random balanced 3:1 partitions based on a repeated Montecarlo sub-sampling ($\frac{3}{4}$ of the data used for training, and $\frac{1}{4}$ for testing).

For our experiments, we considered two populations of classifiers. Population *A* was defined by 30 near-optimal classifiers. Population *B* was defined by 60 classifiers of non-homogeneous performance (30 near-optimal classifiers taken from population *A*, and 30 sub-optimal classifiers). Both populations were composed of neural networks¹ (NN) and fuzzy inference systems² (FIS). In what follows we briefly describe the generation of classifiers.

NN: The variability in the NN classifiers comes from the number of examples for weight updating, the epoch number, and the value of momentum. Networks of 3 layers trained with backpropagation algorithm with blocks of 1–10 examples for weight updating are used, the momentum is randomly selected from [0,0.9] interval, and the epoch number is set from [1,6000] interval.

Near-optimal classifiers adjust the epoch number according to the given momentum while the sub-optimal ones use a random epoch number which is likely to produce over or under fitting.

FIS: The variability in the FIS classifiers comes from the number of terms in the partition for each of the input variables, the method used for designing the corresponding fuzzy sets and the method used for rule induction.

The number of terms is randomly taken in the interval [1,5]. Three methods are used to design the partitions according to a given number of terms: hierarchical fuzzy partitioning [5], regular partitioning, or *K*-means algorithm. Three algorithms are available for rule induction: fast prototyping algorithm [4], Wang and Mendel [17] or fuzzy decision trees [8].

Near-optimal classifiers adjust the number of terms in the partition for each input variable in order to minimize the generalization error, while sub-optimal ones do not. They use the randomly chosen one.

Finally, the SFI knowledge was characterized by λ -measures. As a result, the SFI training started evaluating the fuzzy densities (per classes) from the dataset *Z* as follows:

$$g_j^i = P(z_k \in w_j/f_i^j = \max\{f_i(z_k)\}) - P(z_k \notin w_j/f_i^j = \max\{f_i(z_k)\}) \quad (6)$$

Being $P(z_k \in w_j/f_i^j = \max\{f_i(z_k)\})$ the proportion of correct classification in the class, and $P(z_k \notin w_j/f_i^j = \max\{f_i(z_k)\})$ the “false ones” in the others.

3.1. Datasets

In what follows we briefly describe the datasets used for our experiments.

Table 1
Description of UCI and Grape datasets.

Dataset	Samples	#Attributes	#Classes
Car	1728	6	4
Glass	214	10	6
Iris	150	4	3
Pima	768	8	2
Wine	178	13	3
Yeast	1484	8	10
Grape	400	8	8

Table 2

The mean test errors of the selection-SFI combination rule (*sSFI*) and the best classifier (X_b) are shown along with their mean agreement (*Agr*) between predictions and their mean difference (*Dif*) between mean test errors. *CI* is the 95% confidence interval on difference between mean test errors; Q_r is the mean size of the subset of classifiers induced by the selection-SFI combination rule.

Datasets	<i>sSFI</i>	Q_r	X_b	<i>Agr</i>	<i>Dif</i>	<i>CI</i>
Car	0.0391	5.7	0.0571	0.9523	-0.018	[-0.0234, -0.0127]
Glass	0.0907	4.7	0.1	0.9277	-0.0092	[-0.0277, 0.0092]
Iris	0.0395	4	0.05	0.9789	-0.0105	[-0.0237, 0]
Pima	0.2271	4.5	0.2302	0.9021	-0.0031	[-0.0151, 0.0083]
Wine	0.0222	5.2	0.0267	0.9955	-0.0044	[-0.0089, 0]
Yeast	0.4204	3.6	0.4334	0.9180	-0.0129	[-0.0205, -0.0054]
Grape	0.236	4.5	0.28	0.818	-0.044	[-0.066, -0.022]

Benchmark datasets: Table 1 shows the characteristics of six datasets from UCI³ repository.

Real dataset (Grapes): The objective of *grape* problem, the data are provided by Cemagref, is to determine their variety from an external analysis done by near infrared spectrum method of 512 wavelengths. Depending on their physical meaning, experts select 8 wavelengths to constitute the input variables of classifiers. The dataset consists of 50 examples for each grape variety. The output space is composed of 8 classes: carignan, grenache blanc, chardonnay, roussane, marselan, mourvèdre, grenache noir and clairette.

3.2. Results and comparisons

Considering the aim of multiclassification, the first comparison of the proposed selection-SFI combination approach was made against the best classifier (X_b). The second one considered a popular untrained combination rule such as the majority voting (MV). Finally, our method was compared against the state of art selection approach developed by [3] which relies on the clusterization of classifiers (implemented by *hclust* function of *stats* package, R⁴ library). Methods based on exhaustive analysis [14] were not considered owing to their high computational costs; the number of possible subsets to be evaluated equals $\sum_{i=1}^n \binom{n}{i}$, *n* arbitrarily large. For similar reasons, we did not consider the full *SFI* combination of classifiers: even using a simplified model of measures (λ -measures), the hard root finding of *n* – 1-order polynomial is required. Finally, regarding the evaluation of the statistical significance of the differences between test errors, a bootstrap approach with a confidence interval of 95% (*p*-value 0.05) was used [12].

Table 2 shows the performance of the selection-SFI combination rule and the best classifier on populations *A* and *B*. Since population *B* is derived from population *A* by the addition of sub-optimal classifiers, the identity of the best classifier remains unchanged. Remarkably, a similar behavior was observed on the proposed method: the classification performance, the size, and the composition of the selected subset of classifiers remains unchanged on

¹ <http://fuzzy.cs.uni-magdeburg.de/~borgelt/software.html>.

² <http://www.inra.fr/Internet/Departements/MIA/M/fispro/>.

³ <http://www.ics.uci.edu/~mllearn/MLRepository.html>.

⁴ <http://www.r-project.org/>.

Table 3

The mean test errors of the selection-SFI combination (*sSFI*) and the majority voting (*MV*) rules on population *A* are shown along with their mean agreement (*Agr*) between predictions and their mean difference (*Dif*) between mean test errors. *CI* is the 95% confidence interval on difference between mean test errors; Q_r is the mean size of the subset of classifiers induced by the selection-SFI combination rule.

Datasets	<i>sSFI</i>	Q_r	<i>MV</i>	<i>Agr</i>	<i>Dif</i>	<i>CI</i>
Car	0.039	5.7	0.06	0.954	-0.021	[-0.026, -0.015]
Glass	0.091	4.7	0.152	0.854	-0.061	[-0.087, -0.035]
Iris	0.039	4	0.042	0.982	-0.003	[-0.013, 0.008]
Pima	0.227	4.5	0.222	0.932	0.005	[-0.005, 0.0146]
Wine	0.022	5.2	0.04	0.982	-0.018	[-0.029, -0.009]
Yeast	0.420	3.6	0.411	0.889	0.009	[-0.0003, 0.018]
Grape	0.236	4.5	0.289	0.841	-0.053	[-0.073, -0.033]

Table 4

The mean test errors of the selection-SFI combination (*sSFI*) and the majority voting (*MV*) rules on population *B* are shown along with their mean agreement (*Agr*) between predictions and their mean difference (*Dif*) between mean test errors. *CI* is the 95% confidence interval on difference between mean test errors; Q_r is the mean size of the subset of classifiers induced by the selection-SFI combination rule.

Datasets	<i>sSFI</i>	Q_r	<i>MV</i>	<i>Agr</i>	<i>Dif</i>	<i>CI</i>
Car	0.039	5.7	0.141	0.877	-0.102	[-0.111, -0.094]
Glass	0.091	4.7	0.224	0.815	-0.133	[-0.161, -0.104]
Iris	0.039	4	0.05	0.974	0.0105	[-0.024, 0.003]
Pima	0.227	4.5	0.3	0.785	-0.073	[-0.090, -0.056]
Wine	0.022	5.2	0.067	0.947	-0.044	[-0.062, -0.029]
Yeast	0.420	3.6	0.602	0.642	-0.182	[-0.197, -0.166]
Grape	0.236	4.5	0.349	0.797	-0.113	[-0.135, -0.09]

population *B*, which suggests that the selection-SFI combination rule is robust w.r.t the addition of sub-optimal classifiers. In addition, whatever the size of the underlying population of classifiers, rather small subsets of classifiers were induced by the selection-SFI combination: an average of 4 classifiers was selected across all datasets, which is easily manageable by the SFI combination rule. Finally, the selection-SFI combination rule performed equal or better than the best classifier on both populations (p -value 0.05).

The performance of the selection-SFI combination and the majority voting rules on populations *A* and *B* are respectively shown in Tables 3 and 4. We should note that majority voting is degraded on population *B* due to the presence of sub-optimal classifiers. On the other hand, the power of the selection-SFI combination rule remains unchanged.

Finally, the performance of the selection-SFI combination and the clusterization-selection [3] rules is shown in Tables 5 and 6 for populations *A* and *B*, respectively. Similarly to the majority voting rule, clusterization-selection is slightly degraded by the presence of sub-optimal classifiers. In addition, it tends to select larger subsets of classifiers; a possible explanation for these results

Table 5

The mean test errors of the selection-SFI combination (*sSFI*) and the clusterization-selection (*CS*) rules on population *A* are shown along with their mean agreement (*Agr*) between predictions and their mean difference (*Dif*) between mean test errors. *CI* is the 95% confidence interval on difference between mean test errors; Q_r and N_{cs} are the mean size of the subset of classifiers induced by the selection-SFI combination and the clusterization-selection rules, respectively.

Dataset	<i>sSFI</i>	Q_r	<i>CS</i>	N_{cs}	<i>Agr</i>	<i>Dif</i>	<i>CI</i>
Car	0.039	5.7	0.055	19	0.959	-0.016	[-0.021, -0.011]
Glass	0.091	4.7	0.117	29	0.911	-0.025	[-0.046, -0.005]
Iris	0.039	4	0.042	29	0.981	-0.003	[-0.013, 0.008]
Pima	0.227	4.5	0.226	10	0.927	0.001	[-0.009, 0.011]
Wine	0.022	5.2	0.031	17	0.987	-0.009	[-0.018, 0]
Yeast	0.420	3.6	0.429	4.5	0.897	-0.008	[-0.017, 0]
Grape	0.236	4.5	0.239	3.9	0.827	-0.003	[-0.025, 0.018]

Table 6

The mean test errors of the selection-SFI combination (*sSFI*) and the clusterization-selection (*CS*) rules on population *B* are shown along with their mean agreement (*Agr*) between predictions and their mean difference (*Dif*) between mean test errors. *CI* is the 95% confidence interval on difference between mean test errors; Q_r is the mean size of the subset of classifiers induced by the selection-SFI combination rule.

Dataset	<i>sSFI</i>	Q_r	<i>CS</i>	N_{cs}	<i>Agr</i>	<i>Dif</i>	<i>CI</i>
Car	0.039	5.7	0.053	15.6	0.962	-0.019	[-0.019, -0.01]
Glass	0.091	4.7	0.126	18.7	0.909	-0.035	[-0.055, -0.015]
Iris	0.039	4	0.039	37.5	0.968	0	[-0.016, 0.016]
Pima	0.227	4.5	0.222	18.2	0.935	-0.005	[-0.005, 0.015]
Wine	0.022	5.2	0.04	47.4	0.969	-0.018	[-0.031, -0.004]
Yeast	0.420	3.6	0.444	3	0.856	-0.023	[-0.034, -0.013]
Grape	0.236	4.5	0.251	24.6	0.827	-0.015	[-0.036, -0.006]

is that the clusterization mechanism may be puzzled by the presence of a rather large proportion of sub-optimal classifiers (half of the members of population *B* are sub-optimal classifiers).

Overall, experimental results suggest that the proposed approach can be useful to boost the combination of arbitrary sets of classifiers. This hypothesis was verified on the Grapes dataset, for which further screening revealed that combination improvements were achieved from individuals with a large proportion of errors. In other words, the proposed method was able to exploit the complementary distribution of errors.

4. Conclusions

We considered the problem of efficient and effective decision-making from an arbitrary population of classifiers. A selection-combination approach based on the Sugeno fuzzy integral was proposed. The requirement of efficiency was accomplished by means of a greedy algorithm designed for the identification of prospective subsets of classifiers under the Sugeno fuzzy integral. The requirement of effectiveness was attained with a heuristic search that takes into account the fuzzy integral behavior. Experimental results on benchmark and real datasets showed that the performance of proposed selection-combination is at least compatible with those of the best, majority vote, or clusterization and selection classifier, which suggests their practical usefulness for identifying multiclassifiers from arbitrary populations of classifiers.

References

- [1] G. Brassard, P. Bratley, Fundamentals of Algorithmics, Prentice-Hall, 1996.
- [2] G. Brown, J. Wyatt, R. Harris, X. Yao, Diversity creation methods: a survey and categorisation, Information Fusion, Elsevier Pub. 6(1) (2005) 5–20.
- [3] G. Giacinto, F. Roli, Design of effective neural network ensembles for image classification purposes, Image and Vision Computing 19 (9) (2001) 699–707.
- [4] P. Glorionec, Algorithmes d' apprentissage pour systèmes d' inférence floue, Editions Hermès, Paris, 1999.
- [5] S. Guillaume, B. Charnomordic, Generating an interpretable family of fuzzy partitions, IEEE Transactions on Fuzzy Systems 12 (3) (2004) 324–335.
- [6] L. Hansen, P. Salamon, Neural network ensembles, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (1990) 993–1001.
- [7] Tin Kam Ho, Multiple classifier combination: lessons and next steps, Hybrid Methods in Pattern Recognition 74 (1) (2002) 171–198.
- [8] H. Ichihashi, T. Shirai, K. Nagasaka, T. Miyoshi, Neuro-fuzzy ID3: a method of inducing fuzzy decision trees with linear programming for maximizing entropy and an algebraic method for incremental learning, Fuzzy Sets and Systems 81 (1996) 157–167.
- [9] L. Kuncheva, C. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Machine Learning 51 (2) (2002) 181–207.
- [10] T. Murofushi, M. Sugeno, Fuzzy measures, fuzzy integrals, Fuzzy measures and integrals, in: M. Grabisch, T. Murofushi, M. Sugeno (Eds.), Theory and Applications, Physica-Verlag, Heidelberg, 2000, pp. 3–41.
- [11] D. Partridge, W. Yates, Engineering multiversion neural-net systems, Neural Computation 8 (4) (1996) 869–893.

- [12] R. Rifkin, A. Klautau, In defense of one-vs-all classification, *Machine Learning Research* 5 (2004) 101–141.
- [13] F. Roli, G. Giacinto, Design of multiple classifier systems, in: H. Bunke, A. Kandel (Eds.), *Hybrid Methods in Pattern Recognition*, World Scientific Publ. Co., 2002, pp. 199–226.
- [14] A. Sharkey, N. Sharkey, U. Gerecke, G. Chandroth, The “test and select” approach to ensemble combination, *Lecture Notes in Computer Science* 1857 (2000) 30–44.
- [15] M. Sugeno, *Theory of fuzzy integrals and its applications*, Ph.D. Thesis, Tokio Institute of Technology, 1974.
- [16] G. Valentini, F. Masulli, Ensembles of learning machines, *Neural Nets WIRN Vietri-02, Series Lecture Notes in Computer Sciences*, Springer-Verlag, Heidelberg, Germany 2486 (2002) 3–19.
- [17] L. Wang, J. Mendel, Generating fuzzy rules by learning from examples, *IEEE Transactions on Systems Man and Cybernetics* 22 (6) (1992) 1414–1427.
- [18] L. Xu, A. Krzyzak, C. Suen, Methods of combining multiple classifiers and their applications to hand-written character recognition, *IEEE Transactions on Systems Man and Cybernetics* 22 (3) (1992) 418–435.