

# Scalable Fuzzy Multiclassifier System

**Pilar Bulacio, Luis Magdalena**  
 ETSI Telecomunicación  
 Universidad Politécnica de Madrid  
 Spain  
 {pb,llayos}@mat.upm.es

**Elizabeth Tapia**  
 Escuela de Ingeniería Electrónica  
 Universidad Nacional de Rosario  
 Argentine  
 etapia@eie.fceia.unr.edu.ar

## Abstract

This paper deals with the design of scalable, fuzzy adaptive, decision making systems constructed from sets of heterogeneous classifiers. We propose a multiclassifier architecture formed by sparsely connected coalitions of classifiers. Coalitions are defined by fuzzy integral operators on small, but not necessarily disjoint, subsets of classifiers. The small size constraint on individual coalitions is intended for both the interpretability of fuzzy measures and low complexity of fuzzy integral operators. In addition, the sparse connection constraint guarantees realizable good independent coalitions amenable to further information fusion stages. Simple rules regarding the number of coalitions, the number of classifiers per coalition and the number of coalitions where each classifier should participate are presented. Experimental results show the feasibility of our proposal.

**Keywords:** Scalability, multiclassifier architecture, fuzzy integral, mutual information, sparse matrix.

## 1 Introduction

The design of multiclassifiers systems is a promising line of research in the field of decision making. Roughly speaking, the design of these systems can be separated in two parts [14]. The first part concerns to the design of the base of classifiers (types,

characteristics, quantity...) and is highly problem specific. The second part, which is common to several applications, is devoted to the design of combination methods over individual classification results.

It is natural that if the number of heterogeneous classifiers ( $n$ ) is increased, the probability of common misclassifications should diminish. In fact, two conditions increase decision making performance [6, 8]: the accuracy of individual information sources (classifiers) and their diversity. However, under an increasing number of classifiers, many combinations methods might yield to inconsistent results and/or increase their computational complexity unacceptably as in the case of the fuzzy integral (FI) approach. Inconsistencies might be due to flaws in information fusion schemes. Particularly, precision problems arise at the estimation of FI parameters when decomposable measures,  $\lambda$ -measures, are used. Regarding the consistency of combining schemes, it seems better to work with small sets of classifiers. Furthermore, many small subsets sharing the property of local diversity can be devised given a sufficient large number of heterogeneous classifiers.

From the above discussion, it follows that a scalable multiclassifier design could be constructed from the rearrangement of classifiers which enhance their complementary and aggregated behavior, according to classifiers features and aggregation operator constraints. In other words, small, almost disjoint, subsets of diverse classifiers should be used.

The paper is organized as follow. In Section 2, we present the problem and the proposed archi-

ture. Specifically, we divide the multiclassifier system study in three stages: the training of the base of classifiers, the estimation of FI parameters and the design of the combinational structure (sizing and selection), which are described in the sections 3 and 4. In Section 3, we introduce FI operators and analyze consistency issues. In Section 4, we go into details about the design of combination architecture, according to classifiers features and FI parameters. In Section 5, we show the whole process in a well-known practical problem. Finally, Section 6 offers some conclusions.

## 2 Problem Statement

Cooperative behavior among classifiers takes place when they have a complementary knowledge. In order to make possible the complementarity among classifiers, we propose the generation of diversity by means of the heterogeneity of the individual decision approaches.

As we said earlier, when the number of quite complementary classifiers is increased, the common misclassification may diminish. But also the combinational methods may become complex or inconsistent. Let us first analyze complexity issues regarding the use of FI operator. FI provides a useful, but extremely constrained, theoretical framework for the fuzzy adaptive multiclassification. Because of the use of fuzzy measures for knowledge characterization, FI combination may include different types of uncertainty with a great power of description, but the estimation of  $2^n - 1$  parameters for  $n$  classifiers also implies a great complexity.

The inconsistency factor is due to blemishes in the underlying information fusion scheme and the method used for the parameters estimation. In fact, to elude the estimation complexity Sugeno [12] introduced the decomposable  $\lambda$ -fuzzy measure, the normalized fuzzy measure with the  $\lambda$ -additivity. But  $\lambda$  is a root of a polynomial whose coefficients are the product of estimated density measures, therefore, its calculation propagates all estimation errors.

Taking into account the above mentioned, we split the base of classifiers in chunks, small enough to avoid knowledge inconsistencies and complexity

but large enough to cover all the problem. Thus, we are now faced with the following two combinatorial problems:

**Problem 1:** Given  $n$  heterogenous classifiers, over which we make  $m$  fuzzy integrals, coalitions of  $v$  classifiers ( $v \ll n$ ), with at most  $t$  times that each classifier can be reused in different coalitions ( $t \ll m$ ), which is the most promising architecture under a weighted majority rule?

**Problem 2:** Furthermore, which are the general rules underlying the reasonable choice of parameters  $n$ ,  $m$ ,  $v$  and  $t$ ?

The process of constructing scalable multiclassifiers is carried out in the following stages:

1. The training of heterogeneous classifiers and the definition of a common framework to manage the uncertainty of different inference methods [11] (probability, possibility...).
2. The estimation of FI parameters. They will be used for both the information fusion and combinational architecture design, estimating the classifiers coverage on the classes space.
3. The design of combinational structure. It consists on the structure dimensioning study, where the sizes of  $m$ ,  $t$ ,  $v$  are estimated, and the structure selection, where a suitable combinational structure among the potential ones is selected, according to the statistical dependencies between classifiers, the coverage and the performance of its coalitions.

We will return to the stages 3 and 4 later when we discuss the decision integration with FI and the process of structure design respectively.

## 3 Decision Integration with FI

Fuzzy integrals have been shown to be an useful method for combining results of multiple sources of information. Its definition with respect to a *fuzzy measure* [12] or *capacity* [2] provides a good framework to represent the imprecise knowledge associated with classifiers. In the literature, practical implementations [1, 5] only combine 2 or 3

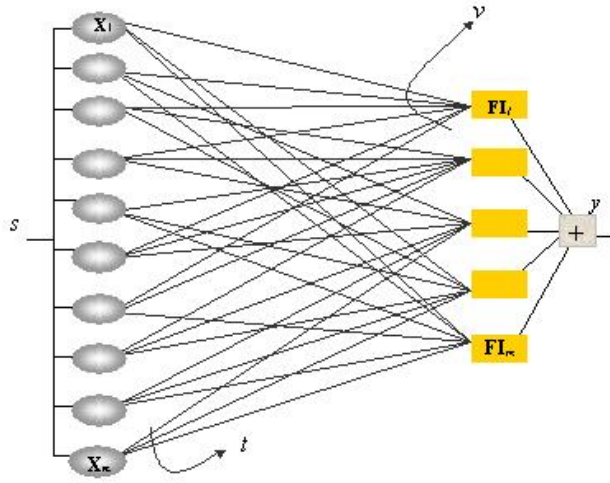


Figure 1: Scalable multiclassifier architecture

classifiers due to the constraint of the parameter estimation with a large number of classifiers.

We will consider the discrete Choquet and Sugeno FIs, viewed as functions of information sources aggregation. See [3, 9] for theoretical details.

Given a multiclassifier system composed of  $n$  classifiers,  $\{X_1, \dots, X_n\}$ , and  $W = \{w_1, \dots, w_C\}$  the set of possible alternatives for a sample  $s \in \mathfrak{R}^p$ , we call  $f(X_i(s))$  or  $f_i : \mathfrak{R}^p \rightarrow [0, 1]^{|W|}$ , the classification function that supplies the individual decision through a vector of  $C$  components. Each component of this vector represents the degree of support given by the classifier  $X_i$  to the hypothesis that  $s$  comes from one class of  $W$ . Then, the multiclassifier decision is obtained by aggregating all partial classifiers evidences weighed by ability degrees ( $g$ ).

Let  $g$  be a fuzzy measure on  $X$ , whose elements are denoted  $X_1, \dots, X_n$ . The Sugeno integral [12] of a function  $f: X \rightarrow [0, 1]$  with respect to  $g$  is defined by

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

The (discrete) Choquet integral [2] of a function  $f : X \rightarrow [0, 1]$  with respect to  $g$  is defined by

$$C_g(f) := \sum_{i=1}^n (f(X_{(i)}) - f(X_{(i-1)}))g(A_{(i)}) \quad (2)$$

Where  $\cdot_{(i)}$  indicates the indices permutation  $0 \leq f(X_{(1)}) \leq \dots \leq f(X_{(n)}) \leq 1$ ,  $f(X_{(0)}) := 0$ , and  $A_{(i)} := \{X_{(i)}, \dots, X_{(n)}\}$ .

A fuzzy measure or capacity  $g$  on  $X$  is a function  $g : 2^X \rightarrow [0, 1]$  if

1.  $g(\emptyset) = 0$ ,
2.  $g(X) = 1$ ,
3.  $A \subset B \subset X \Rightarrow g(A) \leq g(B)$ .

In our case,  $g(A_{(i)})$  quantify the goodness or ability of  $A_{(i)}$  to classify the input on the  $W$  space. In particular, when  $g$  is related to a single element,  $X_i$  with  $i \in \{1, \dots, n\}$ , is called *fuzzy density* of the  $i$ th source or  $g^i$ . The  $g^i$  estimation can be done in different ways, e.g., using probabilities of misclassification [3, 10], through iterative algorithms to diminish the quadratic error [3, 4] or using genetic algorithms [13]. In general, they can not be evaluated from densities.

The  $\lambda$ -fuzzy measure is a decomposable measures proposed by Sugeno to make possible the estimation of fuzzy measures from fuzzy density. They fulfil the further property:

$$g_\lambda^{A,B} = g_\lambda(A \cup B) = g_\lambda(A) + g_\lambda(B) + \lambda g_\lambda(A)g_\lambda(B)$$

for every disjoint subsets  $A, B$  of  $X$  and for some  $\lambda \in (-1, \infty)$ . The value of  $\lambda$  can be uniquely determined for a finite set  $X$  using  $g(X) = 1$ , which leads to solving a polynomial of  $(n-1)$ th degree. The  $\lambda$ -measure simplification reduces to  $n-1$  the number of parameters to be estimated but also limits the flexibility in fuzzy measures relationship.

Dealing with large groups of classifiers would imply a difficulty in the polynomial root determination, but we use small coalitions to attend the precision constraint:  $\lambda$  value propagates the estimation errors of all fuzzy densities. In fact, measures inconsistencies appears before than complexity when  $\lambda$ -measures are used.

#### 4 Process of the Structure Design

In this section we give a perspective of the overall process. We have seen that with a great number of dissimilar classifiers the diversity in misclassifications is possible.

Let us assume  $n$  trained classifiers in both sides: classification and combination. That means, their knowledge bases generated and their fuzzy densities estimated. As we said in the Section 2, we will consider two main objectives: the design of

the combinational structure and the statement of general design rules. We start with the sizing of combination structure parameters and after that the selection of the most data adaptive structure.

#### 4.1 Estimation of Parameters of S

The combinational structure  $S$  consists on  $m$  sparsely connected coalitions of classifiers. To avoid the evidence conflicts or the misinterpretation of the joint knowledge, we would need a structure with small coalitions. Besides, we would require a controlled overlapping among coalitions to obtain relatively independent coalitions results. These features can be fulfilled by *sparse matrices* of  $m \times n$  order. That is, matrices composed of elements  $(s_{i,j}; i=1, \dots, m; j=1, \dots, n)$  taking 1 or 0 values with density (amount of 1s) less than 0.5, randomly generated, which follow these constraints:  $t$  1s per column, times that a classifier is reused and  $v$  1s per row, quantity of classifiers per coalition.

$$S = \begin{bmatrix} 1_{1,1} & \dots & \dots & 0_{1,n} \\ \dots & & & \\ 0_{i,1} & \dots & 1_{i,j} & \dots & 1_{i,n} \\ \dots & & & & \\ 1_{m,1} & \dots & \dots & \dots & 0_{m,n} \end{bmatrix}$$

$s_{i,j}=1$  means the classifier  $j$  takes part in the coalition  $i$ . See [7] for a theoretical background.

The estimation of parameters are mainly associated with the characteristics of the classifiers base, the constraints of the fuzzy measure estimation and the performance required. The parameters are the following:

- Columns: They are defined by the amount of classifiers. The  $n$  value could be evaluated from the expected accuracy, combinational method constraints, classifiers accuracy and classifiers diversity, but in this approach we assume a given  $n$ .
- Number  $v$ : The information fusion has sense if there is redundancy, if not it is only classifiers selection. Therefore, the coalitions size, 1s per row, has a lower bound given by the coalition coverage on  $W$  without redundancy and an upper bound given by the combina-

tion method complexity and the joint knowledge consistency.

- Number  $t$  and  $m$ : The last two parameters,  $t$  (1s per column) and  $m$  (rows or coalitions quantity), have to be defined jointly since the amount of 1s in  $S$  is:  $m \times v = t \times n$ . The  $t$  value measures the overlapping of classifiers in different coalitions. It has an upper bound given by coalitions dependencies and a lower bound to make feasible the local diversity in coalitions.

The next stage is the selection of the most suitable combinational structure according to the features of the set of classifiers.

#### 4.2 Choosing Data Adaptive Structure

The considered features to select an optimal structure are the coverage, performance and mutual information. The *coverage* quantifies, using  $g$ , the system ability over the classes space. It is measured at both the structure and the coalition level, quantifying the coverage on  $W$  of the component coalitions and component classifiers respectively. The *performance* evaluates the coalitions errors using training data. Finally, the *mutual information* (MI) weighs up the dependence among outputs, taking into account the marginal and joint probability distributions of classifiers outputs.

These properties are complementary: the levels of coverage and coalitions performance do not say anything about common fails or similarity among classifiers, and the mutual information says how different classifiers are, but nothing about the classifications goodness.

The *structure selection process* starts from fuzzy densities,  $N$  samples for training the decision combination, and a large number of  $S(n,m,v,t)$  structures. It is done through the following steps:

1. **Coverage indexes of S:** For each  $S$ , it is built a *maximum ability matrix*,  $m \times C$ . Rows are associated with the maximum values of ability for classifying of each coalition. One row is filled with the maximum  $g^i$ , for each class, among the classifiers of one coalition.

Coverage indexes are calculated from mean and standard deviation values of this matrix per classes.

2. **Performance indexes of S:** They are evaluated estimating the performance of each coalition in  $S$  using the data training.
3. **Reordering and first selection:** According to coverage and performance indexes, a group of structures with the highest mean and the lowest standard deviation are selected.
4. **Estimating the MI:** Over the subset of selected structures in the step before, it is estimated the MI per coalitions using the  $N$  training samples. The MI takes into account the marginal and joint probability distributions of classifiers conclusions, quantifying the shared information among them. To do so, the classifiers outputs are discretized within a precision interval  $\text{bin}(j)$ . Being  $c_k^{(i)}$  the  $k^{\text{th}}$  output (classifier) of the  $i^{\text{th}}$  sample ( $i \in [1, N]$ ),  $c_{kj}$  is defined as the number of  $c_k^{(i)}$  values falling in the interval  $\text{bin}(j)$ :

$$c_{kj} = |\{c_k^{(i)} \in \text{bin}(j)\}| ; p(c_{kj}) = \frac{|\{c_k^{(i)} \in \text{bin}(j)\}|}{N}$$

The discrete joint probability mass among  $c_{kj}, c_{k'j'}$  is:

$$p(c_{kj}, c_{k'j'}) = \frac{|\{(c_k^{(i)} \in \text{bin}(j)) \wedge (c_{k'}^{(i)} \in \text{bin}(j'))\}|}{N}$$

And the discrete joint probability mass function among the  $N$  outputs of  $v$  classifiers is:

$$p(c_{1j_1}, \dots, c_{vj_v}) = \frac{|\{\bigwedge_{1 \leq u \leq v} (c_u^{(i)} \in \text{bin}(j_u))\}|}{N}$$

where  $j_u \in \{1, \dots, b\}$ . So, the MI among the  $N$  outputs of  $v$  classifiers is defined as:

$$I(c_1, \dots, c_v) = \sum_{j_1=1}^b \dots \sum_{j_v=1}^b \log \frac{p(c_{1j_1}, \dots, c_{vj_v})}{p(c_{1j_1}) \dots p(c_{vj_v})} \quad (3)$$

For further details see [8].

5. **Reordering and final selection:** According to the MI evaluation, the  $S$  with the highest mean and lowest standard deviation of MI is selected.

## 5 Illustrative Example

We will illustrate an application of scalable multiclassifiers architecture with *wine*<sup>1</sup> data sets.

### A. Data Processing

The *wine* data set contains 178 samples, 13 continuous input variables, 3 output classes. The protocol applied to data set does samplings to generate 50 random sets of 35% samples to train classifiers, 35% to estimate FI parameters, and 30% for system testing.

### B. S parameters

*Columns:*  $n = 10$ , given classifiers with 20% of mean errors. Their diversity is achieved implementing different inference methods (neural networks<sup>2</sup>, clustering and fuzzy inference<sup>3</sup>).

*Coalition dimension:*  $v=3$  estimated from the minimum amount of classifiers to cover all classes within an arbitrary threshold,  $D_{min}=\{1, 2\}$  classifiers needed with  $g_{th}=0.8$ .

*Overlapped:* Due to the fact that we implement only three AI technics, and  $v=3$ , we allow an overlapping or reused classifiers in different coalitions of  $t = \{1; 2\}$ ; to achieve diversity within coalitions, but taking care of the independence among them. Then, with  $v=3, t = \{1; 2\}$ , and  $n = 10$  we may select  $m=7$ , uneven because the combination among rows is using majority voting. It should be noted that the generation of sparse matrices<sup>4</sup> allows to change  $t$  or  $m$  while the other parameters are kept fixed. The results are summarized in the table 1. We noted that using all the classifiers in

Table 1: Scalable Multiclassification Performance

S(n=10,v=3,t={2;3},m=9)	Sugeno	Choquet
Median Error %	2.3	2.3
Mean Error %	3.2	3
Deviation Error %	1.7	1.7
S(n=10,v=10,m=1) Mean Error	~13	~12

the combination the mean error was more than 12% and with the approach proposed the mean error using Sugeno FI was 3.2% or 3% with Choquet FI, better than the best classifier and better than the combination of the complete ensemble.

<sup>1</sup><http://www.ics.uci.edu/~mllearn/MLRepository.html>

<sup>2</sup><http://fuzzy.cs.uni-magdeburg.de/~borgelt/software.html>

<sup>3</sup><http://www.inra.fr/bia/M/fispro/>

<sup>4</sup><http://www.inference.phy.cam.ac.uk/mackay/>

## 6 Conclusions

We have presented a method to make the scalability in trained multiclassifier systems possible. The decision making through a large base of classifiers can be done under an appropriate classifiers reorganization, in coalitions, and a proper selection of the most data-adaptive combinational structure. A suitable reorganization of classifiers, allowing a controlled redundancy, together with measures of the structure coverage, performance and mutual information may provide a framework for solving a general scalability problem.

Attempting to answer the problem 2 (Section 2), for a given  $n$ , the  $v$  value depends on the cardinality of coalitions that can cover the problem and the combination method constraints. The  $t$  and  $m$  values are associated with the overlapping between coalitions which is bounded by dependencies among coalitions and the local diversity of coalitions.

Experimental results show that a given group of diverse classifiers can improve their individual performance and the global performance with a cautious splitting and selection.

### Acknowledgments

The authors would like to thank C. Borgelt, S. Guillaume, D. MacKay, F. Massulli and G. Valentini for publishing their own software tools as free software.

### References

- [1] S. Cho and J. Kim., *Multiple network fusion using fuzzy logic*, Neural Networks, IEEE Transactions on **6(2)** (1995), 497–501.
- [2] G. Choquet, *Theory of capacities*, Annales de l'Institut Fourier **5** (1953), 131–295.
- [3] M. Grabisch, *Fuzzy measures and integrals for decision making and pattern recognition*, TATRA MOUNTAINS Mathematical Publications **13** (1997), 7–34.
- [4] J.M. Keller and J. Osborn, *Training the fuzzy integral*, Int. J. Approx. Reasoning **15** (1996), no. 1, 1–24.
- [5] L.I. Kuncheva, *'fuzzy' vs 'non-fuzzy' in combining classifiers: an experimental study*, Proc LFA'01, Belgium, 2001, pp. 11–22.
- [6] L.I. Kuncheva and C.J. Whitaker, *Ten measures of diversity in classifier ensembles: limits for two classifiers*, In Proc. IEE Workshop on Intelligent Sensor Processing, Birmingham, IEEE., 2001, pp. 10/1–10/6.
- [7] D. MacKay, *Good error correcting codes based on very sparse matrices*, IEEE Transactions on Information Theory **45** (1999), no. 2, 399–431.
- [8] F. Masulli and G. Valentini, *Quantitative evaluation of dependence among outputs in ecoc classifiers using mutual information based measures*, Proceedings of the IJCNN'01, K. Marko and P. Webos (eds.), IEEE, Piscataway, NJ, USA **2** (2001).
- [9] T. Murofushi and M. Sugeno, *Fuzzy measures and integrals: Theory and applications*, ch. Fuzzy Measures and Fuzzy Integrals, pp. 3–41, Physica-Verlag, 2000.
- [10] Tuan D. Pham and Hong Yan, *Information fusion by fuzzy integral*, Intelligent Information Systems (Narasimhan and Jain, eds.), Australian and New Zealand Conference on, Narasimhan and Jain, 1996, pp. 191–194.
- [11] J.J. Sudano, *Pignistic probability transforms for mixes of low- and high-probability events*, International Conference on Information Fusion, Canada **TUB3** (2001), 23–27.
- [12] M. Sugeno, *Theory of fuzzy integrals and its applications*, Ph.D. thesis, Tokio Institute of Technology, 1974.
- [13] D. Wang, X. Wang, and J.M. Keller, *Determining fuzzy integral densities using a genetic algorithm for pattern recognition*, Fuzzy Information Processing Society (1997), 263–267.
- [14] L. Xu, A. Krzyzak, and C. Y. Suen, *Methods of combining multiple classifiers and their applications to hand-written character recognition*, IEEE trans. on Systems, Man and Cybernetics **22(3)** (1992), 418–435.