Sensitive, Specific, and Interpretable: Evolving a Fuzzy Mammographic-Interpretation Assessment Tool

Carlos Andrés Peña-Reyes[†], Moshe Sipper[‡], and Luis Prieto[§]
[†]Swiss Federal Institute of Technology in Lausanne, Switzerland. Email:Carlos.Pena@epfl.ch
[‡]Ben-Gurion University, Beer-Sheva, Israel.
[§]Duran i Reynals Hospital, Barcelona, Spain.

Abstract - In this paper we focus on the problem of mammographic interpretation and in the construction of a computerized system to assess physicians in this task. We propose a solution based on fuzzy systems, and apply our fuzzy modeling methodology—Fuzzy CoCo—to find systems exhibiting both good performance and high interpretability. The results presented show the capacity of Fuzzy CoCo to find systems that satisfy different requirements concerning the complexity of the solution, expressed mainly as the number of rules.

I. Introduction

A major class of problems in medical science involves the definition of treatments and procedures, based upon the interpretation of one or various tests performed upon the patient. When several tests are involved, the ultimate decision may be difficult to obtain, even for a medical expert. This has given rise, over the past few decades, to computerized tools, intended to aid the physician in making sense out of the welter of data. A prime target for such computerized tools is in the domain of cancer detection. Specifically, where breast cancer is concerned, the treating physician is interested in ascertaining whether the patient under examination exhibits the symptoms of a benign case, or whether her case requires further examination. Mammography remains the principal technique for detecting breast cancer. Its undoubtable value in reducing mortality notwithstanding, mammography's positive predictive value (PPV) is low: only between 15 and 35% of mammographic-detected lesions are cancerous [3, 7]. The remaining 65 to 85% of biopsies, besides being costly and time-consuming, cause understandable stress on women facing the doubt of cancer. A computer-based tool that assists radiologists during mammographic interpretation would contribute to increasing the PPV of biopsy recommendations.

A good computerized tool for medical decision support should possess two characteristics, which are often in conflict. First, the tool must attain the highest possible *performance*, i.e., detecting as much as possible the malignant cases, while minimizing the number of unnecessary biopsies. Second, it would be highly beneficial for such a system to be human-friendly, exhibiting so-called *interpretability*. This means that the physician is not faced with a black box that simply spouts answers (albeit correct) with no explanation; rather, we would like for the system to provide some insight as to *how* it derives its outputs.

In this paper we propose to use fuzzy logic-based systems exhibiting the two aforementioned characteristics, to construct a computerized tool to assist mammographic interpretation. To explore the design space of these fuzzy systems, we apply Fuzzy CoCo—a fuzzy modeling technique based on cooperative coevolution conceived to provide a good balance between accuracy and interpretability.

The rest of this paper is organized as follows: In the next section we present briefly Fuzzy CoCo, our cooperative coevolutionary approach to fuzzy modeling. In Section III we describe the mammography interpretation problem, which is the focus of our interest herein. Section IV then describes the application of Fuzzy CoCo to this problem, followed by the results obtained, presented in Section V. Finally, we conclude in Section VI.

II. Fuzzy CoCo: A Cooperative Coevolutionary Approach to Fuzzy Modeling

Fuzzy CoCo is a Cooperative Coevolutionary approach to fuzzy modeling, wherein two coevolving species are defined: database (membership functions) and rule base [4]. This approach is based primarily on the framework defined by Potter [8].

In Fuzzy CoCo, the fuzzy modeling problem is solved by two coevolving cooperative species. Individuals of the first species encode values which define completely all the membership functions for all the variables of the system. Individuals of the second species define a set of rules of the form:

if (v_1 **is** A_1) **and** ... (v_n **is** A_n) **then** ($output$ **is** C),

where the term A_v indicates which one of the linguistic labels of fuzzy variable v is used by the rule. The two evolutionary algorithms used to control the evolution of the two populations are instances of a simple genetic algorithm. The genetic algorithms apply fitness-proportionate selection to choose the mating pool, and apply an elitist strategy with an elitism rate Er to allow a given proportion of the best individuals to survive into the next generation. Standard crossover and mutation operators are applied with probabilities P_c and P_m, respectively.

An individual undergoing fitness evaluation establishes co-

0-7803-7280-8/02/$10.00 ©2002 IEEE

TABLE I
VARIABLES CORRESPONDING TO A PATIENT'S CLINICAL DATA.

v_1	Age	[28-82] years	
v_2	Menstrual history	1	Premenopausal
		2	Postmenopausal
v_3	Family history	1	None
		2	Second familiar
		3	First familiar
		4	Contralateral
		5	Homolateral

operations with one or more representatives of the other species, i.e., it is combined with individuals from the other species to construct fuzzy systems. The fitness value assigned to the individual depends on the performance of the fuzzy systems it participated in. Representatives, or *cooperators*, are selected both fitness-proportionally and randomly from the last generation since they have already been assigned a fitness value. In Fuzzy CoCo, N_{cf} cooperators are probabilistically selected according to their fitness, usually the fittest individuals, thus favoring the exploitation of known good solutions. The other N_{cr} cooperators are selected randomly from the population to represent the diversity of the species, maintaining in this way exploration of the search space. For a more detailed exposition of Fuzzy CoCo see [4].

III. The mammography interpretation problem

The *Catalonia Mammography Database*, which is the object of our study, was collected at the Duran y Reynals hospital in Barcelona. It consists of 15 input attributes and a diagnostic result indicating whether or not a carcinoma was detected after a biopsy. The 15 input attributes include three clinical characteristics (Table I) and two groups of six radiologic features, according to the type of lesion found in the mammography: mass or microcalcifications (Table II).

A radiologist fills out a reading form for each mammography, assigning values for the clinical characteristics and for one of the groups of radiologic features. Then, the radiologist interprets the case using a five-point scale: (1) benign; (2) probably benign; (3) indeterminate; (4) probably malignant; (5) malignant. According to this interpretation a decision is made on whether to practice a biopsy on the patient or not. The Catalonia database contains data corresponding to 227 cases. Each case is examined by three different readers—for a total of 681 readings—but only diverging readings are kept. The actual number of readings in the database is 516, among which 187 are positive (malignant) cases and 329 are negative (benign) cases.

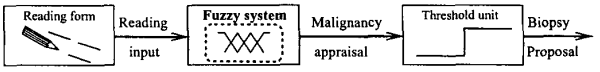

Fig. 1. Proposed system.

IV. An evolved fuzzy system to assess mammographic interpretation

A. Proposed solution

The solution scheme we propose to construct the mammographic-interpretation assessment tool is depicted in Figure 1. It consists of a reading form, a fuzzy subsystem and a threshold unit. Based on the 15 input attributes collected with the reading form, the fuzzy system computes a continuous appraisal value of the malignancy of a case. The threshold unit then outputs a biopsy recommendation according to the fuzzy system's output. The threshold value used in this system is 3, which corresponds to the "indeterminate" diagnostic. Fuzzy CoCo is applied to design the fuzzy system in charge of appraising malignancy.

B. Fuzzy-parameter setup

We used prior knowledge about the Catalonia database to guide our choice of fuzzy parameters. In addition, we took into account the following five semantic criteria, defining constraints on the membership-function parameters [5,6]: (1) distinguishability; (2) justifiable number of elements; (3) coverage; (4) normalization; and (5) complementarity, as well as the following three syntactic criteria, constraining the encoding of the rules [2,5]: (1) completeness; (2) simplicity; and (3) readability. Following the parameter classification presented in [4], we delineate below the fuzzy system's set-up:

- Logical parameters: singleton-type fuzzy systems; min-max fuzzy operators; orthogonal, trapezoidal input membership functions; weighted-average defuzzification.
- Structural parameters: two input membership functions (*Low* and *High*; two output singletons (*benign* and *malignant*); a user-configurable number of rules. The relevant variables are one of Fuzzy CoCo's evolutionary objectives.
- Connective parameters: the antecedents and the consequent of the rules are searched by Fuzzy CoCo. The algorithm also searches for the consequent of the default rule. All rules have unitary weight.
- Operational parameters: the input membership-function values are to be found by Fuzzy CoCo. For the output singletons we used the values 1 and 5, for *benign* and *malignant*, respectively.

TABLE II

VARIABLES CORRESPONDING TO RADIOLOGIC FEATURES. THERE ARE TWO GROUPS OF VARIABLES USED TO DESCRIBE THE MAMMOGRAPHIC EXISTENCE OF MICROCALCIFICATIONS AND MASSES (LEFT AND RIGHT COLUMNS RESPECTIVELY).

Microcalcifications		Mass	
v_4	Disposition	v_{10}	Morphology
1	Round	1	Oval
2	Indefinite	2	Round
3	Triangular or Trapezoidal	3	Lobulated
4	Linear or Ramified	4	Polilobulated
v_5	Other signs of group form	5	Irregular
1	None	v_{11}	Margins
2	Major axis in direction of nipple	1	Well delimited
3	Undulating contour	2	Partially well delimited
4	Both previous	3	Poorly delimited
v_6	Maximum diameter of group	4	Spiculated
	[3-120] mm	v_{12}	Density greater than parenchyma
v_7	Number	1	Not
1	<10	2	Yes
2	10 to 30	v_{13}	Focal distortion
3	>30	1	Not
v_8	Morphology	2	Yes
1	Ring shaped	v_{14}	Focal asymmetry
2	Regular sharp-pointed	1	Not
3	Too small to determine	2	Yes
4	Irregular sharp-pointed	v_{15}	Maximum diameter
5	Vermicular, ramified		[5-80] mm
v_9	Size irregularity		
1	Very regular		
2	Sparingly regular		
3	Very irregular		

C. Genome encodings

Fuzzy CoCo thus searches for four parameters: input membership-function values, relevant input variables, and antecedents and consequents of rules. To encode these parameters into both species' genomes, which together describe an entire fuzzy system, it is necessary to take into account the heterogeneity of the input variables as explained below.

1. Species 1: Membership functions. The fifteen input variables ($v_1 - v_{15}$) present three different types of values: continuous (v_1,v_6, and v_{15}), discrete ($v_3 - v_5$ and $v_7 - v_{11}$), and binary (v_2 and $v_{12} - v_{14}$). It is not necessary to encode membership functions for binary variables as they can only take on two values. The membership-function genome encodes the remaining 11 variables—three continuous and eight discrete—each with two parameters P_1 and P_2, defining the membership-function apices. Table III delineates the parameters encoding the membership-function genome.

TABLE III

GENOME ENCODING OF PARAMETERS FOR MEMBERSHIP-FUNCTION SPECIES. GENOME LENGTH IS 106 BITS.

Variable type	Qty	Params.	Bits	Total bits
Continuous	3	2	7	42
Discrete	8	2	4	64
	Total Genome Length			106

2. Species 2: Rules. The i-th rule has the form: **if** (v_1 **is** A_1^i) **and ... and** (v_{15} **is** A_{15}^i) **then** (*output* **is** C^i), where A_j^i can take on the values: 1 (*Low*), 2 (*High*), or 0 or 3 (*don t-care*). C^i can take on the val-

ues: 1 (*benign*) or 2 (*malignant*). However, as mentioned before, each database case presents three clinical characteristics and six radiologic features according to the type of lesion found: mass or microcalcifications (note that only a few special cases contain data for both groups). To take advantage of this fact, the rule-base genome encodes, for each rule, 11 parameters: the three antecedents of the clinical-data variables, the six antecedents of one radiological-feature group, an extra bit to indicate whether the rule applies for mass or microcalcifications, and the rule consequent. Furthermore, the genome contains an additional parameter corresponding to the consequent of the default rule. Relevant variables are searched for implicitly by allowing the algorithm to choose non-existent membership functions as valid antecedents ($A_j^i = 0$ or $A_j^i = 3$); in such case the respective variable is considered irrelevant, and removed from the rule. Table IV delineates the parameters encoding the rules genome.

TABLE IV
GENOME ENCODING OF PARAMETERS FOR RULES SPECIES. GENOME LENGTH IS $20 \times N_r + 1$ BITS, WHERE N_r DENOTES THE NUMBER OF RULES.

Parameters	Qty	Bits	Total bits
Clinical data	$3 \times N_r$	2	$6 \times N_r$
Radiologic data	$6 \times N_r$	2	$12 \times N_r$
Rule-type selector	N_r	1	N_r
Consequents	$N_r + 1$	1	$N_r + 1$
Total Genome Length			$20 \times N_r + 1$

D. Evolutionary parameters

Table V delineates values and ranges of values of the evolutionary parameters. The algorithm terminates when the maximum number of generations, G_{max}, is reached (we set $G_{max} = 700 + 200 \times N_r$, i.e., dependent on the number of rules used in the run), or when the increase in fitness of the best individual over five successive generations falls below a certain threshold (10^{-4} in our experiments). Note that mutation rates are relatively higher than those used with a simple genetic algorithm.

E. Fitness function

As the main function of the proposed system is the assessment of a medical diagnosis, our fitness definition takes into account medical diagnostic criteria. The most commonly employed measures of the validity of diagnostic procedures are the sensitivity and specificity, the likelihood ratios, the predictive values, and the overall classification (accuracy) [1]. Table VI provides expressions for four of these measures which are important for evaluating the performance of our systems. Three of them are used in the fitness function, the last one is used in Section V to support the analysis of the results. Besides these criteria, the fitness function provides extra selective pressure based on two syntactic criteria: simplicity and readability (see [5]).

TABLE V
FUZZY COCO SET-UP.

Parameter	Values
Population size N_p	90
Maximum generations G_{max}	$700 + 200N_r$
Crossover probability P_c	1
Mutation probability P_m	{0.005,0.01}
Elitism rate E_r	{0.1,0.2}
"Fit" cooperators N_{cf}	1
Random cooperators N_{cr}	1

TABLE VI
DIAGNOSTIC PERFORMANCE MEASURES. THE VALUES USED TO COMPUTE THE EXPRESSIONS ARE: TRUE POSITIVE (TP): THE NUMBER OF POSITIVE CASES CORRECTLY DETECTED, TRUE NEGATIVE (TN): THE NUMBER OF NEGATIVE CASES CORRECTLY DETECTED, FALSE POSITIVE (FP): THE NUMBER OF NEGATIVE CASES DIAGNOSED AS POSITIVE, AND FALSE NEGATIVE (FN): THE NUMBER OF POSITIVE CASES DIAGNOSED AS NEGATIVE.

Performance criteria	20-rule
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Positive predictive value (PPV)	$\frac{TP}{TP+FP}$

Our fitness function combines the following five criteria: 1) F_{sens}: sensitivity, or true-positive ratio, computed as the percentage of positive cases correctly classified; 2) F_{spec}: specificity, or true-negative ratio, computed as the percentage of negative cases correctly classified (note that there is usually an important trade-off between sensitivity and specificity which renders difficult the satisfaction of both criteria); 3) F_{acc}: classification performance, computed as the percentage of cases correctly classified; 4) F_r: rule-base size fitness, computed as the percentage of unused rules (i.e., the number of rules that

are never fired and can thus be removed altogether from the system); and 5) F_v: rule-length fitness, computed as the average number of *don t-care* antecedents (i.e., unused variables) per rule.

The fitness function is computed in three steps—basic fitness, accuracy reinforcement, and size reduction—as explained below:

1. Basic fitness. Based on sensitivity and specificity, it is given by

$$F_1 = \frac{F_{sens} + \alpha F_{spec}}{1+\alpha},$$

where the weight factor $\alpha = 0.3$ reflects the greater importance of sensitivity.

2. Accuracy reinforcement. Given by

$$F_2 = \frac{F_1 + \beta F'_{acc}}{1+\beta},$$

where $\beta = 0.01$. $F'_{acc} = F_{acc}$ when $F_{acc} > 0.7$; $F'_{acc} = 0$ elsewhere. This step slightly reinforces the fitness of high-accuracy systems.

3. Size reduction. Based on the size of the fuzzy system, it is given by

$$F = \frac{F_2 + \gamma F_{size}}{1+2\gamma},$$

where $\gamma = 0.01$. $F_{size} = (F_r + F_v)$ if $F_{acc} > 0.7$ and $F_{sens} > 0.98$; $F_{size} = 0$ elsewhere. This step rewards top systems exhibiting a concise rule set, thus directing evolution toward more interpretable systems.

V. Results

This section describes the results obtained when applying the methodology described in Section IV. We first delineate the success statistics relating to the evolutionary algorithm. Then, we present the diagnostic performance of two selected evolved fuzzy systems that exemplify our approach.

A total of 65 evolutionary runs were performed, all of which found systems whose fitness exceeds 0.83. In particular, considering the best individual per run (i.e., the evolved system with the highest fitness value), 42 runs led to a fuzzy system whose fitness exceeds 0.88 , and of these, 6 runs found systems whose fitness exceeds 0.9; these results are summarized in Figure 2.

Table VII shows the results of the best systems obtained. The maximum number of rules per system was fixed at the outset to be between ten and twenty-five.

As mentioned before, our fitness function includes two syntactic criteria to favor the evolution of good diagnostic systems exhibiting interpretable rule bases. Concerning the simplicity of the rule base, rules that are encoded in a genotype but that never fire are removed from the phenotype (the final system), rendering it more interpretable. Moreover, to improve readability, the rules are allowed (and encouraged) to contain *don t-care* conditions. The relatively low values of R_{eff} and V_r in Table VII confirm the reinforced interpretability of the evolved fuzzy systems.

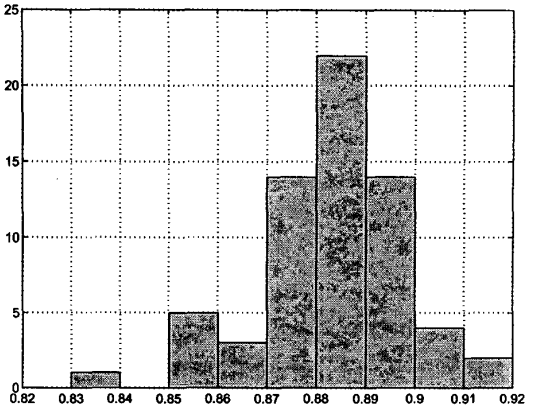

Fig. 2. Summary of results of 65 evolutionary runs. The histogram depicts the number of systems exhibiting a given fitness value at the end of the evolutionary run. The fitness considered is that of the best individual of the run.

TABLE VII

RESULTS OF THE BEST SYSTEMS EVOLVED. RESULTS ARE DIVIDED INTO FOUR CLASSES, IN ACCORDANCE WITH THE MAXIMUM NUMBER OF RULES-PER-SYSTEM, GOING FROM 10-RULE SYSTEMS TO 25-RULE ONES. SHOWN BELOW ARE THE FITNESS VALUES OF THE TOP SYSTEMS AS WELL AS THE AVERAGE FITNESS PER CLASS, ALONG WITH THE NUMBER OF RULES WHICH EFFECTIVELY USED BY THE SYSTEM (R_{eff}) AND THE AVERAGE NUMBER OF VARIABLES PER RULE (V_r).

Max.	Best individual			Average per class		
rules	Fitness	R_{eff}	V_r	Fitness	R_{eff}	V_r
10	0.8910	9	2.22	0.8754	9.17	2.52
15	0.8978	12	2.50	0.8786	12.03	2.62
20	0.9109	17	2.41	0.8934	14.15	2.59
25	0.9154	17	2.70	0.8947	15.78	2.76

Table VIII shows the diagnostic performance measures of two selected evolved systems. The first system, which is the top one over all 65 Fuzzy CoCo runs, is a 17-rule system exhibiting a sensitivity of 99.47% (i.e., it detects all but one of the positive cases), and a specificity of 68.69% (i.e., 226 of the 329 negative cases are correctly detected as benign). The second system is the best found when searching for ten-rule systems. The sensitivity and the specificity of this 9-rule system

TABLE VIII

DIAGNOSTIC PERFORMANCE OF TWO SELECTED EVOLVED SYSTEMS. SHOWN BELOW ARE THE SENSITIVITY, THE SPECIFICITY, THE ACCURACY, AND THE POSITIVE PREDICTIVE VALUE (PPV) OF TWO SELECTED EVOLVED SYSTEMS. IN PARENTHESES ARE THE VALUES, EXPRESSED IN NUMBER OF CASES, LEADING TO SUCH PERFORMANCE MEASURES. THE 17-RULE SYSTEM IS THE TOP SYSTEM. THE 9-RULE SYSTEM IS THE BEST FOUND WHEN SEARCHING FOR TEN-RULE SYSTEMS.

	17-rule	9-rule
Sensitivity	99.47% (186/187)	98.40% (184/187)
Specificity	68.69% (226/329)	64.13% (211/329)
Accuracy	79.84% (412/516)	76.55% (395/516)
PPV	64.36% (186/289)	60.93% (184/302)

are, respectively, 98.40% and 64.13%. As mentioned in Section I, the usual positive predictive value (PPV) of mammography ranges between 15 and 35%. As shown in Table VIII, Fuzzy CoCo increases this value beyond 60%—64.36% for the 17-rule system—while still exhibiting a very high sensitivity.

VI. Concluding Remarks

We presented the application of Fuzzy CoCo, a fuzzy modeling technique based on cooperative coevolution, to the design of a fuzzy mammographic-interpretation assessment tool exhibiting both good performance and high interpretability. In fuzzy modeling, the interpretability-accuracy trade-off is of crucial import, imposing several conditions on the input and output membership functions as well as on the rule definition. As Fuzzy CoCo is highly configurable, it facilitates the management of the mentioned trade-off.

Applying Fuzzy CoCo to breast-cancer diagnosis we concentrated on increasing the interpretability of solutions, obtaining excellent results. We note, however, that the consistency of the entire rule base and its compatibility with the specific domain knowledge can only be assessed by further interaction with medical experts (radiologists, oncologists). Besides, the developed tool must be fine-tuned through further tests submitted to the subjective reading of different radiologists.

References

[1] H. Brenner. Measures of differential diagnostic value of diagnostic procedures. *Journal of Clinical Epidemiology*, 49(12):1435–1439, December 1996.

[2] S. Guillaume. Designing fuzzy inference systems from data: An interpretability-oriented review. *IEEE Transactions on Fuzzy Systems*, 9(3):426–443, June 2001.

[3] S. G. Orel, N. Kay, C. Reynolds, and D. C. Sullivan. Bi-rads categorization as a predictor of malignancy. *Radiology*, 211(3):845–880, June 1999.

[4] C. A. Peña-Reyes and M. Sipper. Fuzzy CoCo: A cooperative-coevolutionary approach to fuzzy modeling. *IEEE Transactions on Fuzzy Systems*, 9(5):727–737, October 2001.

[5] C.-A. Peña-Reyes and M. Sipper. Fuzzy CoCo: Balancing accuracy and interpretability of fuzzy models by means of coevolution. In J. Casillas, O. Cordon, F. Herrera, and L. Magdalena, editors, *Trade-off between Accuracy and Interpretability in Fuzzy Rule-Based Modelling*, Studies in Fuzziness and Soft Computing. Physica-Verlag, 2002. (Submitted).

[6] W. Pedrycz and J. Valente de Oliveira. Optimization of fuzzy models. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, 26(4):627–636, August 1996.

[7] L. Porta, R. Villa, E. Andia, and E. Valderrama. Infraclinic breast carcinoma: Application of neural networks techniques for the indication of radioguided biopsias. In J. Mira, R. Moreno-Díaz, and J. Cabestany, editors, *Biological and Arti`cial Computation: From Neuroscience to Technology*, volume 1240 of *Lecture Notes in Computer Science*, pages 978–985. Springer, 1997.

[8] M. A. Potter and K. A. DeJong. Cooperative coevolution: An architecture for evolving coadapted subcomponents. *Evolutionary Computation*, 8(1):1–29, spring 2000.